Environmental Quality and Education

Senioritis Members: Katelyn Greene, Anisha Kumar, Vanessa Serrano, Kana Yamamoto

Introduction

Environmental justice issues explore the discrepancies in the impact of environmental degradation felt by different cultural and socioeconomic groups. Often, those who suffer most from environmental pollution and public health issues are lower income and/or minority groups (Bullard 2000). This is largely attributed to the fact that housing costs tend to be cheaper in environmentally degraded locations, and the low-income, marginalized groups often cannot find a better, affordable choice (Bullard 2000). Moreover, since public education systems are funded through local tax returns, populations in lower income communities also tend to have a lower quality of education (Ward 2006). This contributes to a cycle in which those born into a lower economic standing have less mobility due to poor education and fewer economic opportunities (Ward 2006).

In our project, we analyzed the relationship between environmental health and socioeconomic status in the United States by looking at the influence of household income and various environmental factors on educational attainment. Our primary objective was to see if lower socioeconomic status correlated with higher environmental pollution. To tackle this objective, we collected data on income and education for socioeconomic variables and data on air quality, water quality, available green space and tree canopy for environmental health variables.

Initially, we considered analyzing our data at the state-level, but income and education levels can differ vastly within a state. Generally, income and education tend to be higher near more populated, metropolitan areas and lower in less populated, rural areas. By conducting a county-level analysis rather than a larger, state-level analysis, we minimize the loss of information due to generalization and obtain a more accurate relationship between the variables.

The key environmental factors examined were particulate matter 2.5 (PM 2.5) concentration in the atmosphere to assess air quality, arsenic level in drinking water to assess water quality, percent tree cover and available green space to assess environmental access. Particulate matter is a mixture of microscopic solids and liquid droplets, and is largely associated with industrial and automobile pollution (EPA 2016a). More specifically, PM 2.5 is defined as the concentration of particulate matter greater than 2.5 micrograms per cubic meter and is one of the major causes of smog (EPA 2016a). It is linked to severe health problems affecting the lung and the heart (EPA 2016b). High PM 2.5 concentrations have been known to increase the chances of heart attacks, aggravate asthma, decrease lung function, and even cause premature death in people with existing conditions (EPA 2016b). Arsenic pollution in drinking water is also an important environmental issue and a reliable determinant of water quality. Although arsenic is a naturally occurring element in rocks and minerals (Welch et al. 1999), high concentrations of arsenic ingestion can lead to cancer in the skin, lung, and bladder (NRC 1999). It also causes liver and kidney damage (USGS 2016). Arsenic pollution most commonly results from pesticides, industrial waste, and the extraction of common metals such as copper, lead, and zinc (USGS 2016). Additionally, areas with high geothermal activity will also have higher arsenic contamination (Welch et al. 1999). Aside from environmental pollution, we also looked at the general environmental quality of the community using tree canopy cover and available green space as indicators. Increasing urban green spaces are known to increase quality of life and promote public health, but the lack thereof in many minority and low-income communities has been raised as an environmental justice issue (Wolch et al. 2014).

Note: All the code used for data collection and visualization is located in the appendix and as a live project in this github repo.

Data Gathering

As mentioned previously, we obtained data exclusively at the county-level to ensure there is sufficient data to determine if a correlation exists between environmental quality and socioeconomic factors. It should be mentioned that Alaska, Hawaii and, several US territories were omitted from all datasets to facilitate ease of mapping and create consistency between datasets as many were missing data for these areas. The variables of interest relating to education, income, and environmental quality were obtained via web scraping and source download links. Data cleaning then ensued in order to make it glyph-ready.

County Identifier Data (Appendix: "County FIPS Reference Table Data")

In order to create an aggregate data table that contained all the variables we were interested in, we needed a common identifier for the counties. Since the United States maintains a standardized five number code called a Federal Information Processing Standard (FIPS) code for each county, we used this FIPS Code as the primary key to join all individual data tables. The first two numbers of the FIPS code indicate the state and the following three numbers indicate the county in the state. By using the FIPS Code, we were able to ensure the format of county names was consistent and prevent duplicate entries. As a result, the basis of our aggregate data table was a webscraped table from Wikipedia that contained the county FIPS codes, county names, state FIPS codes, and state names.

Education Data (Appendix: "Educational Attainment Data")

We downloaded our educational attainment data as an excel file from the United States Department of Agriculture (USDA) Economic Research Service (See dataset). This county-level data included the percent of adults who did not have a high school diploma, had a high school diploma, had some college/associate's degree, and those who had a Bachelor's degree and higher. The years included were 1970, 1980, 1990, 2000, and an estimate for 2015 based on the 2010 census. We used the estimate for 2015 because we thought it was best to focus on the most recent data. To obtain the data, we downloaded the excel sheet from the website url, stored it in a temporary file, then read the relevant sheet and row numbers into a data frame.

We then needed to clean the data to isolate the variables of interest: the FIPS code, the percent of adults in the county with a given level of educational attainment, and the Rural-Urban Continuum Code. To clarify, Rural-Urban Continuum code is an integer scale from 1-9 that designates a county as metropolitan, urban, or rural based on population density. Based on the dataset documentation, we classified each county as metropolitan (code 1-3), urban (code 4-7) and rural (code 8-9). While we obtained data relating to several levels of educational attainment, we ultimately chose to only graph our variables against percent of adults with a Bachelor's degree because we thought it would yield more interesting results.

Income Data (Appendix: "Income Data")

Median household income data also came from the USDA datasets (see dataset). We used the same data collection method we used to obtain the education data. The raw county-level data included the county FIPS code, unemployment rate between 2007-2015, median household income in 2015, Urban Influence Code, and Rural-Urban Continuum code. We then extracted the FIPS code and median household income in 2015 for our analysis. Additionally, we decided to divide income into four quartiles so that we could gauge how income related to environmental quality. After dividing the income data into its quartiles, our data fell into the following categories: <$40,000, $40,000-$47,000, $47,000-$54,000, and >$54,000.

Particulate Matter Data (Appendix: "Air Quality Data")

Air quality data came from the Center for Disease Control in units of micrograms per cubic meter (see dataset). The data included various measurement parameters for PM from 2001-2013. Since this data was stored in a standard csv format, we used read.csv() to read in raw data from the download url and store it in a data frame. We first cleaned the data by filtering out all years except 2010 because this year had data for the most counties. We then selected the variable "average measure of particulate matter 2.5 per year based off of monitored/modeled measurements" as it was the most complete category of PM 2.5 measurement. However, this dataset was not updated to reflect county name changes that occurred after 2013, so we manually had to change the FIPS code for Shannon County to the FIPS for the newly named Oglala county in South Dakota.

Tree Canopy Data (Appendix: "Forestry Data")

Urban forest data was collected from the USDA Forest Service (Northern Research station). See link to investigate the dataset. The raw data includes compiled results regarding population, tree canopy cover, and surface cover from forest assessments at the community, county, and state level in 2010.

To conduct our analysis for this project, we needed to access and download the county-level forestry data from this site. However, the website only offers data downloads for each state, thus we conducted webscraping and data wrangling in order to attain a full data set. To create a full data table containing information about all the counties, we scraped the source webpage to find the links to all the states' webpages. Each state's webpage was then scraped to find the state-level data download link in the form of an excel file. Each state's excel file was downloaded using temporary files, cleaned of excess metadata rows, and then read into R. A unique challenge of this dataset was learning to interact with and extract data from excel files with several sheets. We then extracted all the county names and their respective values for tree canopy (m^2/person), available green space (ha), and tree canopy cover in developed regions (%).

While the county names of this dataset were based on counties designated by the US Census Bureau, the data did not include each county's unique FIPS code identifier and the county name data was outdated, only reflecting census data before the 2000's. This prevented us from automatically joining all the county-level data from each state together effectively. As a result, we then mapped all the county names in each state to their respective FIPS codes, using the reference table from the US Census Bureau. We then manually identified county name changes not reflected in the forestry data and updated them automatically to ensure county names to match those in the Census Bureau reference table of FIPS codes. Lastly the county level data frames were joined together to form one large data frame of all US counties, FIPS code identifiers, and forestry variables.

Arsenic Data (Appendix: "Arsenic Data")

The arsenic concentration data was from the United States Geological Survey (USGS) (See dataset). It contained point data of groundwater arsenic concentrations measured at over 20,000 locations across the US from 1973 to 2001. Arsenic has a long residence time in the environment, so we assumed that arsenic concentrations do not change substantially over a 40-year period and thus we used all the available data for the continental US.

This data was stored as an excel file on the USGS water quality data page, so we downloaded the file as a temp file and read it into RStudio. The raw data included 12 variables of which only five were useful. These variables included the state, FIPS code, measured arsenic concentrations (ug/L), and the latitudes and longitudes of where the measurements were taken. After selecting these five variables, we cleaned the data to only include measurements taken in the continental US. Then we checked the structure of the data frame and realized that many of the numeric variables were classified as a character type, and hence we changed them to a numeric type as needed. We also deleted rows that contained missing values for any of the variables because any missing value within a row made the observation unusable.

Once the arsenic data was cleaned, the second step was to assign an arsenic concentration value to each county. Because the arsenic data was not taken for every county, we performed a spatial interpolation with the available point data using an Inverse Distance Weighted (IDW) method from the geospatial package called gstat. An IDW interpolation essentially creates a continuous surface of estimated values based on the given point measurements. It is called IDW because the algorithm to calculate the estimates takes into account how far a point on the surface is to the the point measurements--the estimated value will be more similar to the point measurement if the point of estimation is closer to the point measurement and more different if the two are farther apart. This method also accounts for clusters of points--if a cluster of points has very similar values, a larger area surrounding the cluster will have a similar value, assuming that the data stays relatively consistent. The caveat to this method is that if there is a small area in which arsenic concentrations are extremely high, it may overestimate the surrounding arsenic concentrations. Overall, this method is simply an estimation, but it provides consistent and continuous data that is sufficient within the scope of this project.

Before performing the interpolation, we first had to turn the arsenic data frame into a geospatial object with the same coordinate projections as the US counties map (detailed explanation of the counties map acquisition is addressed in the “Leaflet Maps” section). This conversion was accomplished by assigning the latitude and longitude to a coordinate system that allowed for the point measurements to be spatially located onto a fixed location on the globe. Once the arsenic data was converted into a geospatial object, we ran the IDW interpolation across the boundaries of the continental US. The interpolation estimated the arsenic concentrations for roughly every 0.25 sq-mi area. This created a better estimate for county arsenic levels because it requires an aggregation of data. By aggregating the data to the bounds of each county, we smooth any extreme values. This aggregated estimated mean arsenic concentration for each county was calculated using raster functions from the “raster” package (a raster is a continuous spatial surface). Once we completed the aggregation, we had a geospatial object with an estimated arsenic concentration for each county. Finally, we converted this data back into a regular data frame with only the necessary variables--the FIPS codes and estimated arsenic concentrations.

Compiling the Aggregate Data Frame (Appendix: “Acquire and Join all Data”)

The code we wrote to download and read each data set into R was organized into a hierarchy of separate functions. This allowed us to automatically read and load the individual data frames within one main code chunk by calling all the functions sequentially. We then used a join function from the plyr package to join all individual data tables by county FIPS code and create one large aggregate data frame. This large data frame includes all our variables for each county to facilitate ease of later wrangling and visualization.

Visualization

Scatter Plots/Regressions (Appendix: “Scatter Plots” and “Linear Regressions “)

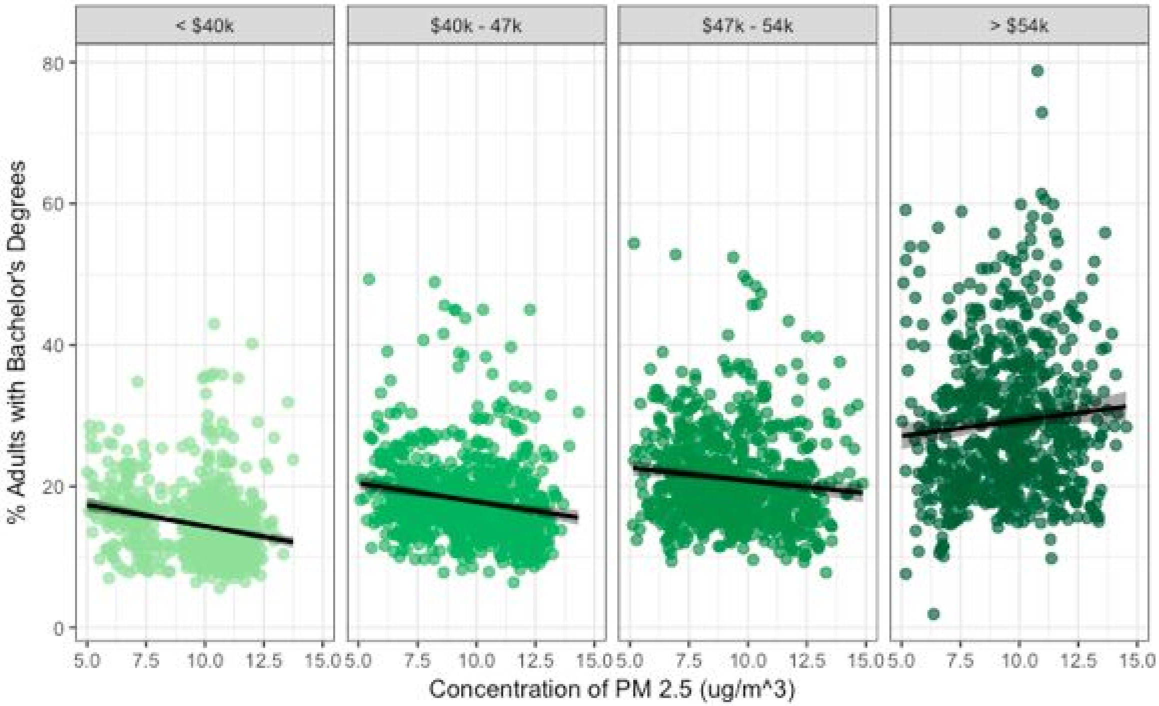


Figure 1 : Particulate Matter vs Education, grouped by Income Quartile

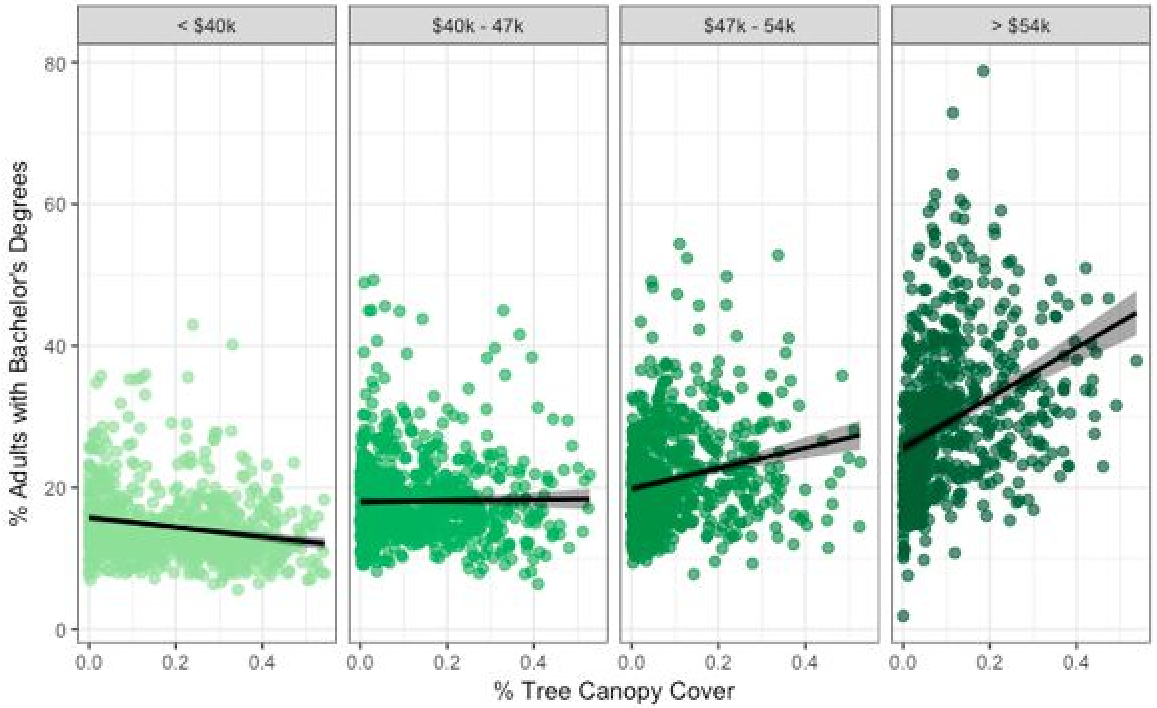


Figure 2: Tree Canopy Cover vs Education, grouped by Income Quartile

Visualizing Environmental Variables and Educational Attainment by Income

We were interested in exploring the relationship between environmental quality factors and education. In order to visualize the relationship, we plotted percent of adults with a Bachelor's degree or higher vs. PM 2.5 concentration, one of our indicators of environmental quality. Originally we intended to look at a ratio of the percent of people with higher education to the percent of people with lower education. However, upon visualization we realized this was an unreliable metric because of the questionable ratio accuracy. Consequently, we instead focused on the percent of adults with Bachelor's degrees.

As shown in **Figure 1**, we chose to facet our graphs by income quartile to explore how income level corresponded to an area's level of pollution and education. Based on the graph, all quartiles except for the highest quartile had a negative slope. This suggests that more air pollution corresponds to lower education level. However, after some speculation, we realized that education levels could be heavily dependent on other confounding variables such as the type of region a person lives (metropolitan, rural vs. urban).

Additionally, we visualized percent of adults with Bachelor's degree vs. tree canopy cover. However, the graph contained outliers that skewed the regression line and is not shown. Thus after further investigation of the data, we felt justified in filtering outliers where tree canopy cover was greater than 55% of the county since these few outliers were not representative of the other thousands of counties. The resulting graph (**Figure 2)** revealed a more consistent spread in the data and exposed an interesting trend where the higher the tree canopy coverage, the more adults with a Bachelor's degree. However, the lowest income quartile in particular caught our attention, as the slope had a negative correlation while the other income quartiles had a positive correlation. This negative correlation among low income counties could be due to the fact that rural areas inherently have a lot of trees and lower education.

Based on our assessment of **Figures 1 and 2**, it became apparent that examining our data based on income quartiles was not sufficient to determine the relationship between environmental quality and education. For this reason, we omitted the graphs of our other two variables, available green space and arsenic pollution faceted by income. Instead, we decided to examine our data in the context of rural-urban classification.

Visualizing Environmental Variables and Educational Attainment by Rural-Urban Classification

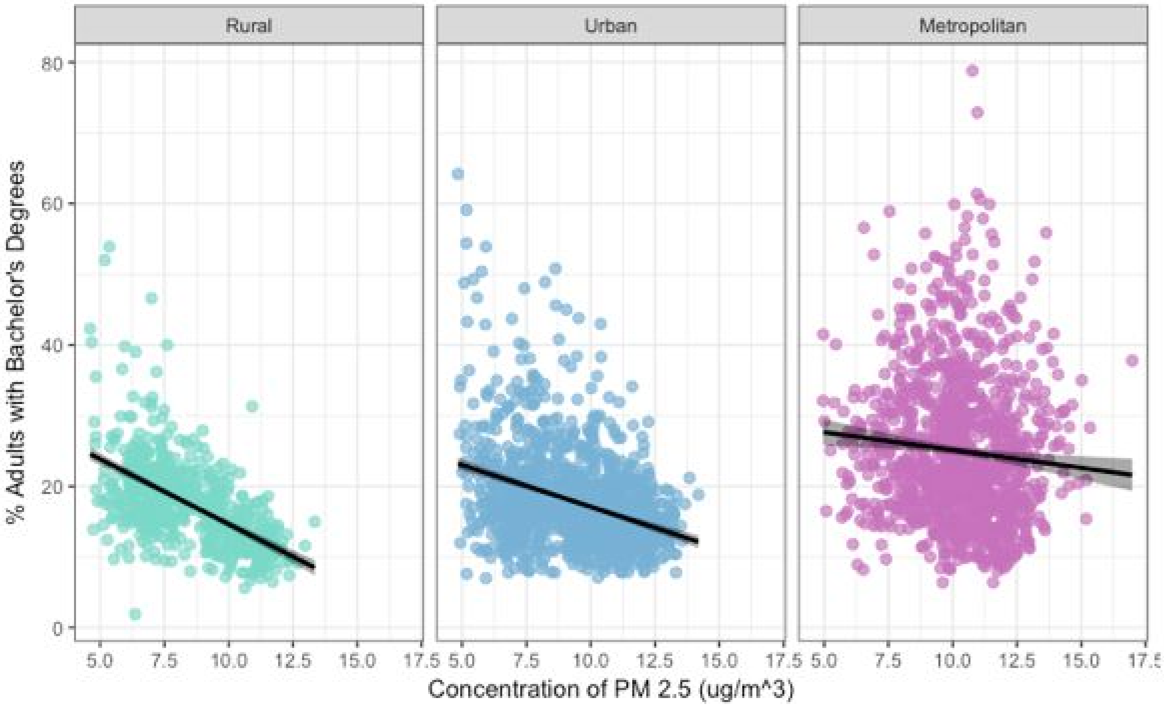
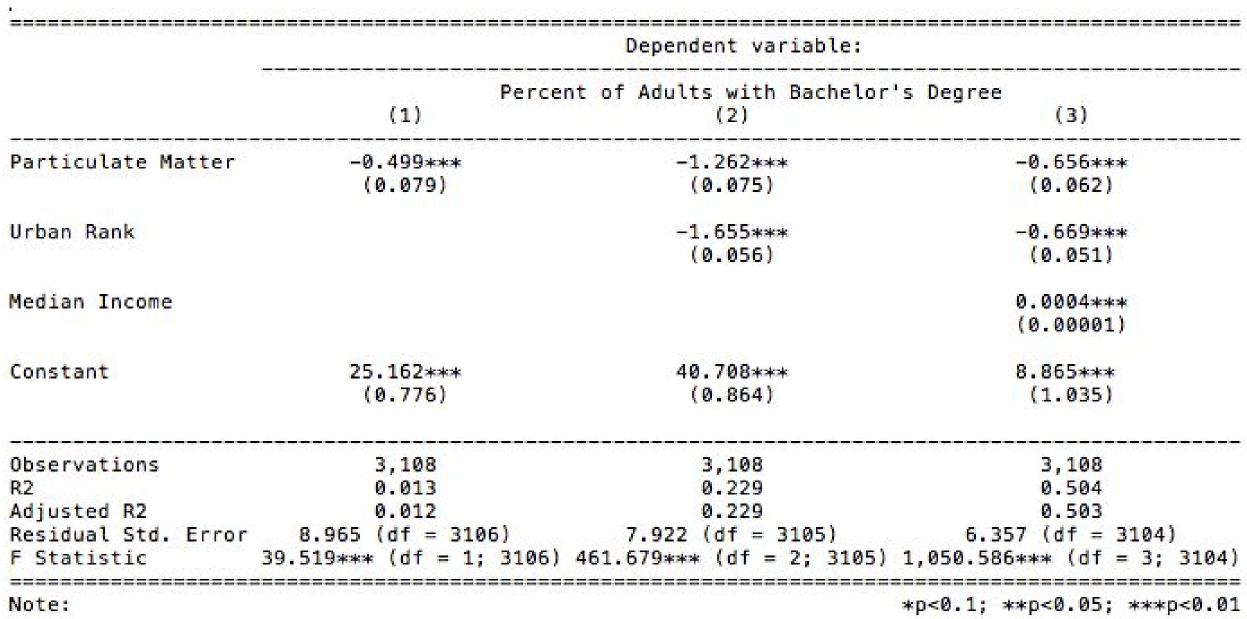


Figure 3: Education vs Particulate Matter, grouped by Rural-Urban Classification

Since a region degree of urbanization (i.e. rural, urban or metropolitan) can also influence the environmental quality, such as air quality, we decided to facet by a variable we created called Urban Rank. Urban Rank uses the Rural-Urban Continuum Code to designate a county as either rural (code 1-3), urban (code 4-7), or metropolitan (code 8-9). We then plotted percent of adults with a Bachelor's degree against air quality (concentration of PM 2.5). As demonstrated in **Figure 3**, all three regions have a negative slope, suggesting that areas with higher PM 2.5 have a lower percentage of adults with higher education. The slope was steepest for counties in rural areas, suggesting that in rural areas there exists a stronger correlation between PM 2.5 and education.



```
==========================================================================================
                                              Dependent variable:
                    ----------------------------------------------------------------------
                                     Percent of Adults with Bachelor's Degree
                             (1)                        (2)                        (3)
------------------------------------------------------------------------------------------
Particulate Matter        -0.499***                  -1.262***                  -0.656***
                           (0.079)                    (0.075)                    (0.062)

Urban Rank                                           -1.655***                  -0.669***
                                                      (0.056)                    (0.051)

Median Income                                                                   0.0004***
                                                                                (0.00001)

Constant                  25.162***                  40.708***                  8.865***
                           (0.776)                    (0.864)                    (1.035)

------------------------------------------------------------------------------------------
Observations                3,108                      3,108                      3,108
R2                          0.013                      0.229                      0.504
Adjusted R2                 0.012                      0.229                      0.503
Residual Std. Error    8.965 (df = 3106)          7.922 (df = 3105)          6.357 (df = 3104)
F Statistic        39.519*** (df = 1; 3106) 461.679*** (df = 2; 3105) 1,050.586*** (df = 3; 3104)
==========================================================================================
Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Figure 4: Regression for Education vs Particulate Matter

To further examine the significance of these results, we ran three linear regressions to test the relationship between the percent of adults and amount of PM 2.5, as shown in **Figure 4**. The first regression compared the two variables of interest, percent of adults with Bachelor's degree and amount of PM 2.5 in a county. The second regression included Urban Rank as a control and the third included both Urban Rank and Income as control variables.

The results were statistically significant for all the regressions we ran. We can reject the null hypothesis that the coefficient of interest is 0, meaning there is some relationship between education and air quality. The R^2 rose for every control variable that we added, but it never rose higher than 0.503. This indicates that there is some variable that we are not accounting for in our regression. Given more time, we would explore additional control variables such as a metric for car use in each county, the number of manufacturing factories in each county, or the demographics of each county.

The correlation was negative for all the regressions we ran, meaning that higher PM 2.5 is associated with a lower percent of people with Bachelor’s degrees. This makes sense because we would expect areas with high amounts of air pollution to have lower educational attainment since PM 2.5 primarily comes from either farming or manufacturing, two industries that are associated with lower educational attainment. However, it is imperative that we remain cautious about drawing definitive conclusions from these regressions.

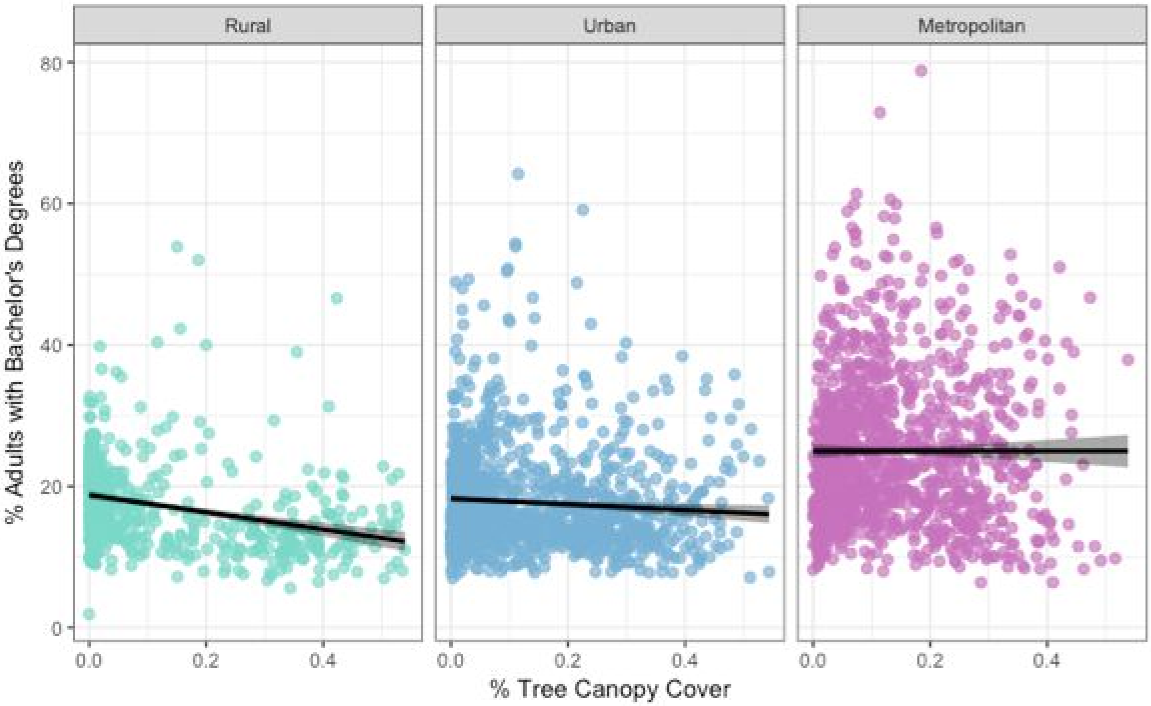
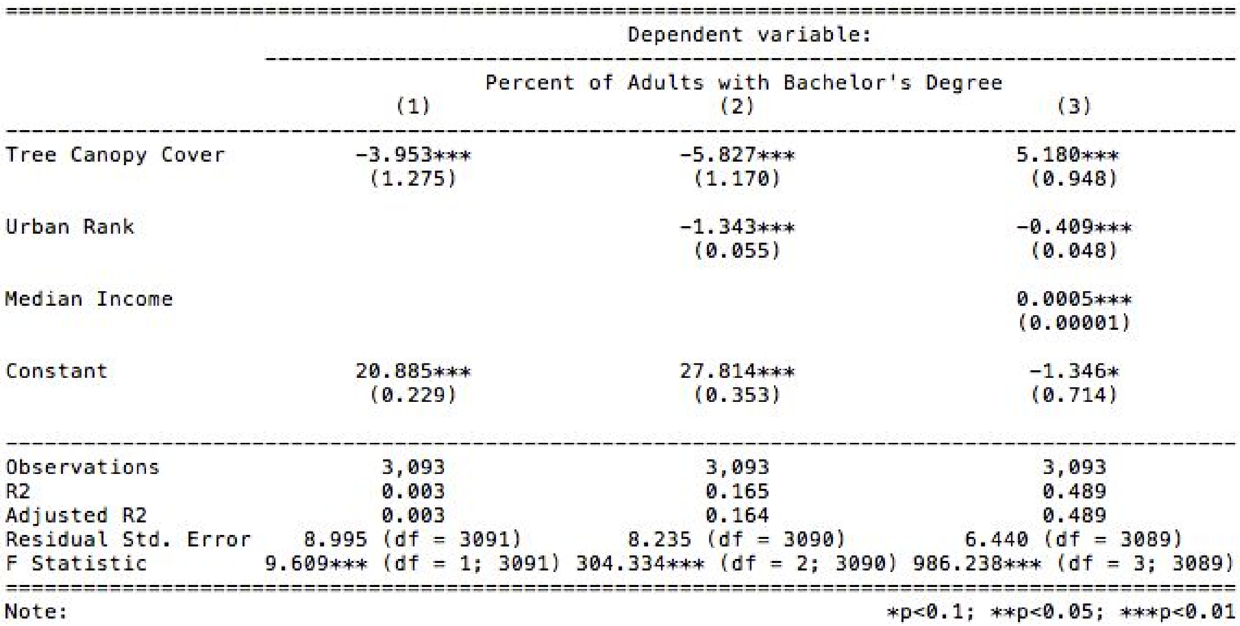


Figure 5: Education vs Tree Canopy Cover, grouped by Rural-Urban Classification

Another way we assessed environmental quality was by collecting data pertaining to the tree canopy cover in each county. Once again we placed percent of adults with Bachelor's degrees or higher on the y-axis and percent tree canopy cover on the the x-axis. As mentioned previously, it is important to categorize counties by their degree of urbanization, especially for a variable such as tree canopy cover since rural regions inherently experience less development and thus more tree cover than urban or metropolitan areas.

As shown in **Figure 5**, the resulting graphs all contain a slightly negative slope. Moreover, the rural counties show the steepest negative slope, indicating that counties with higher percent of tree canopy cover tend to have fewer adults with a Bachelor's degree. Urban regions had a very small correlation with tree canopy cover while metropolitan areas had almost no correlation. This lack of correlation among metropolitan counties is perhaps attributed to the fact that these areas are heavily populated and densely developed and thus have less available land area available for widespread foliage. Since **Figure 5'**s slopes are less steep than those of **Figure 3**, it can be inferred that there is less of a correlation between education and tree canopy cover than between education and concentration of PM 2.5.



```
===================================================================================
                                    Dependent variable:
                   ----------------------------------------------------------------
                             Percent of Adults with Bachelor's Degree
                           (1)                    (2)                   (3)
-----------------------------------------------------------------------------------
Tree Canopy Cover       -3.953***              -5.827***              5.180***
                         (1.275)                (1.170)               (0.948)

Urban Rank                                     -1.343***             -0.409***
                                                (0.055)               (0.048)

Median Income                                                         0.0005***
                                                                     (0.00001)

Constant                20.885***              27.814***             -1.346*
                         (0.229)                (0.353)               (0.714)

-----------------------------------------------------------------------------------
Observations              3,093                  3,093                 3,093
R2                        0.003                  0.165                 0.489
Adjusted R2               0.003                  0.164                 0.489
Residual Std. Error    8.995 (df = 3091)      8.235 (df = 3090)     6.440 (df = 3089)
F Statistic        9.609*** (df = 1; 3091) 304.334*** (df = 2; 3090) 986.238*** (df = 3; 3089)
===================================================================================
Note:                                                     *p<0.1; **p<0.05; ***p<0.01
```

Figure 6: Regression of Education vs Tree Canopy Cover

We then ran three regressions to further investigate the relationship between education and tree canopy cover shown in **Figure 6**. The first regression was between the two variables of interest, percent of adults with Bachelor's and tree canopy cover. The second regression included Urban Rank as a control, and the third included both Urban Rank and income as control variables.

We found that all of our coefficients were statistically significant, so we can reject the null hypothesis that the coefficient of interest is 0. The R^2 increased as we added control variables; however, it never became sufficiently large. Thus, we are likely missing a stronger explanatory variable. For example, it would have been beneficial to see how the relationship would change if we were to add demographic information as a control.

We found negative relationships between educational attainment and tree canopy for the first two regressions, which is expected given the results from the scatter plot. It was interesting that the sign of the coefficient changed from positive to negative when we included the income control variable. Additionally, this sign change caused the R^2 to make a relatively large jump from 0.165 to 0.489. This leads us to believe that the relationship is actually positive and not negative, meaning that a higher percent of tree canopy cover in a county is actually associated with a higher percent of adults with Bachelor's degrees, contradicting what we saw in the scatterplots. However, we need to be careful about interpreting these results. They do not explicitly suggest a causal relationship and at best suggest there is a positive correlation between percent of adults with Bachelor’s degrees and acres of tree canopy cover.

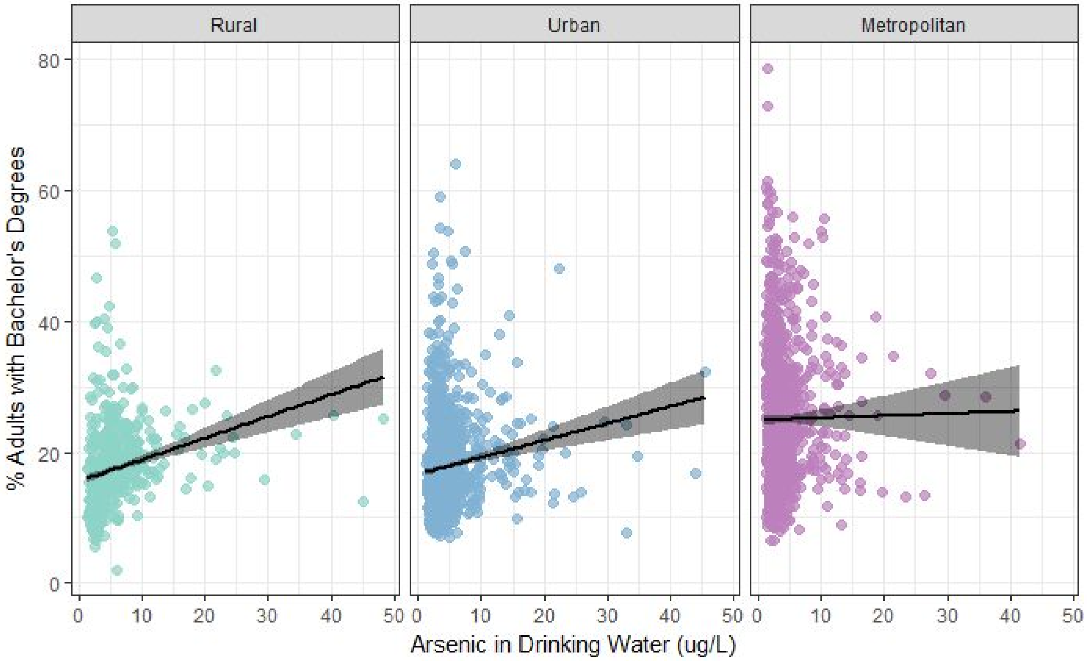
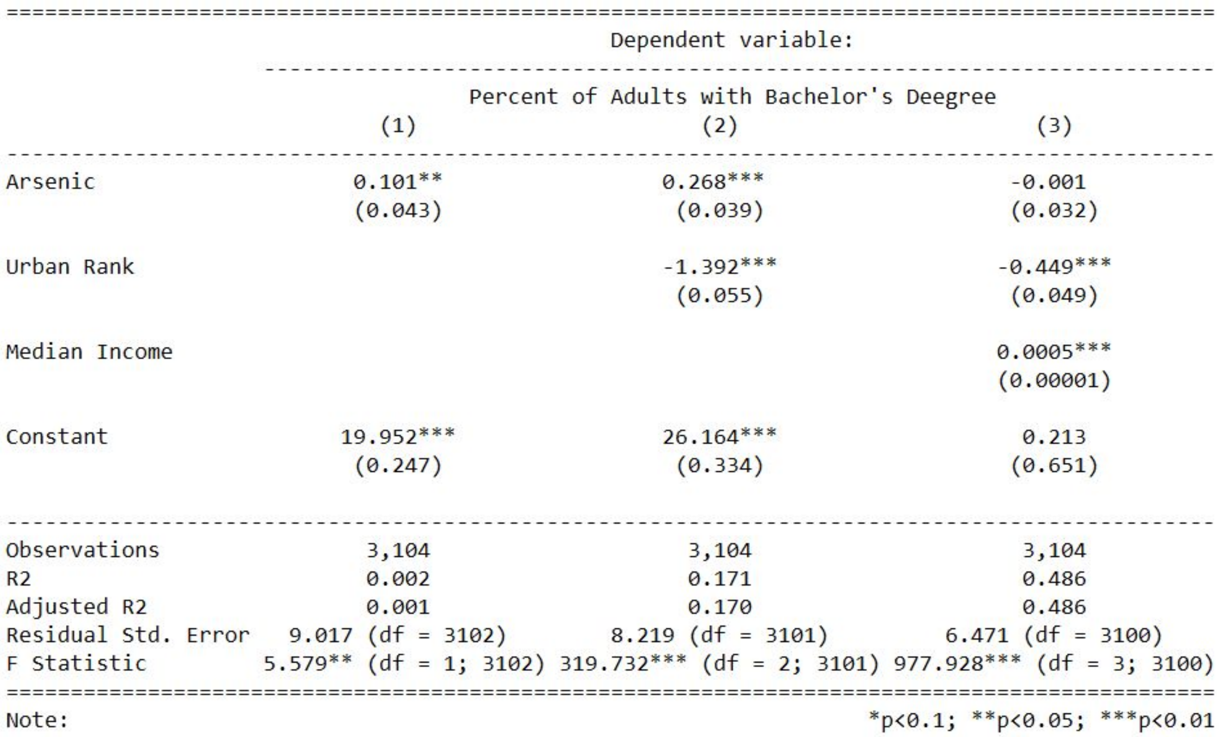


Figure 7: Education vs Arsenic, grouped by rural-urban classification

Another variable we included to assess environmental quality was arsenic content in water. Before performing our analysis, we used preliminary visualizations to identify outliers and omit arsenic concentrations greater than 50 ug/L (detailed explanation for this cut-off is provided in the section on geospatial mapping of arsenic data). Once again we placed percent of adults with Bachelor's degrees or higher on the y-axis and percent tree canopy cover on the the x-axis. To reiterate, it is important for us to categorize counties by their degree of urbanization especially for a variable such as tree canopy cover, since rural regions would innately contain more undeveloped land than urban or metropolitan areas.

As shown in **Figure 7**, the resulting graphs have a slightly positive slope. The rural region has the steepest positive slope, indicating that the more micrograms of arsenic there are per liter of water, the more the percent of adults with a Bachelor's degree. Urban regions had a very small correlation with arsenic content while metropolitan areas had almost no correlation.



```
=================================================================================
                                     Dependent variable:
                     ------------------------------------------------------------
                             Percent of Adults with Bachelor's Deegree
                              (1)                   (2)                    (3)
---------------------------------------------------------------------------------
Arsenic                     0.101**              0.268***                 -0.001
                           (0.043)                (0.039)                (0.032)

Urban Rank                                      -1.392***               -0.449***
                                                 (0.055)                (0.049)

Median Income                                                           0.0005***
                                                                        (0.00001)

Constant                   19.952***             26.164***                0.213
                           (0.247)                (0.334)                (0.651)

---------------------------------------------------------------------------------
Observations                3,104                  3,104                  3,104
R2                          0.002                  0.171                  0.486
Adjusted R2                 0.001                  0.170                  0.486
Residual Std. Error   9.017 (df = 3102)       8.219 (df = 3101)     6.471 (df = 3100)
F Statistic       5.579** (df = 1; 3102) 319.732*** (df = 2; 3101) 977.928*** (df = 3; 3100)
=================================================================================
Note:                                                 *p<0.1; **p<0.05; ***p<0.01
```

Figure 8: Regression for Education vs Arsenic

We then ran three regressions to test the relationship between education and the concentration of arsenic in local water, as seen in **Figure 8**. The first regression included only the two variables of interest, percent of adults with Bachelor's and micrograms of arsenic per liter of water in county. The second regression included Urban Rank as a control and the third included both Urban Rank and income as control variables.

The results were statistically significant for all the regressions we ran, so we can reject the null hypothesis that the coefficient of interest is 0. The R^2 rose for every control variable that we added, but it never rose higher than 0.408. Similar to our other analyses, this indicates that there is some variable that we are not accounting for in our regression. Given more time, we would include a metric for the number of mines in each county or the demographics of each county.

The correlation was positive for the first two regressions involving arsenic, but negative for the last. We intuitively expected a negative correlation, i.e. higher arsenic content would be associated with lower percent of people with Bachelor's degrees. However, only our last regression, where we controlled for income, showed this result. We don't have a robust explanation the positive correlation between arsenic and educational attainment. We can extrapolate that the correlation may have changed signs in the third regression because it controls for income, and people with higher incomes don't typically live in areas with high arsenic content in water. Ultimately though, we must be careful about drawing any conclusions without more rigorous statistical analysis .

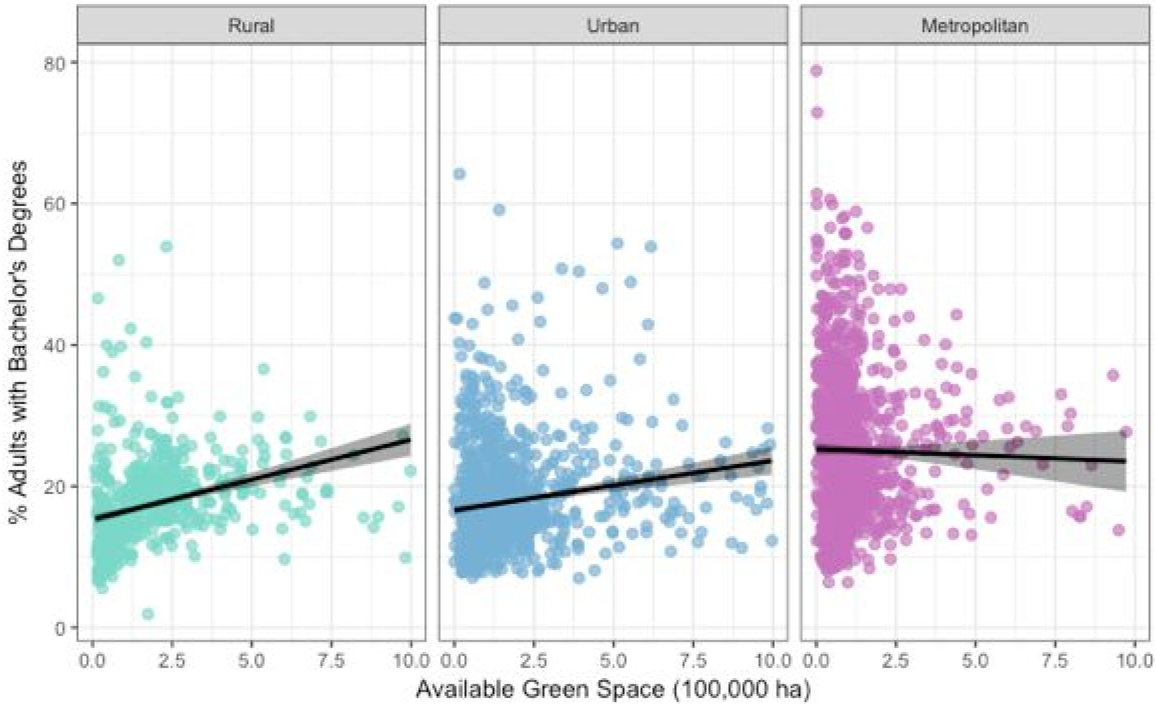
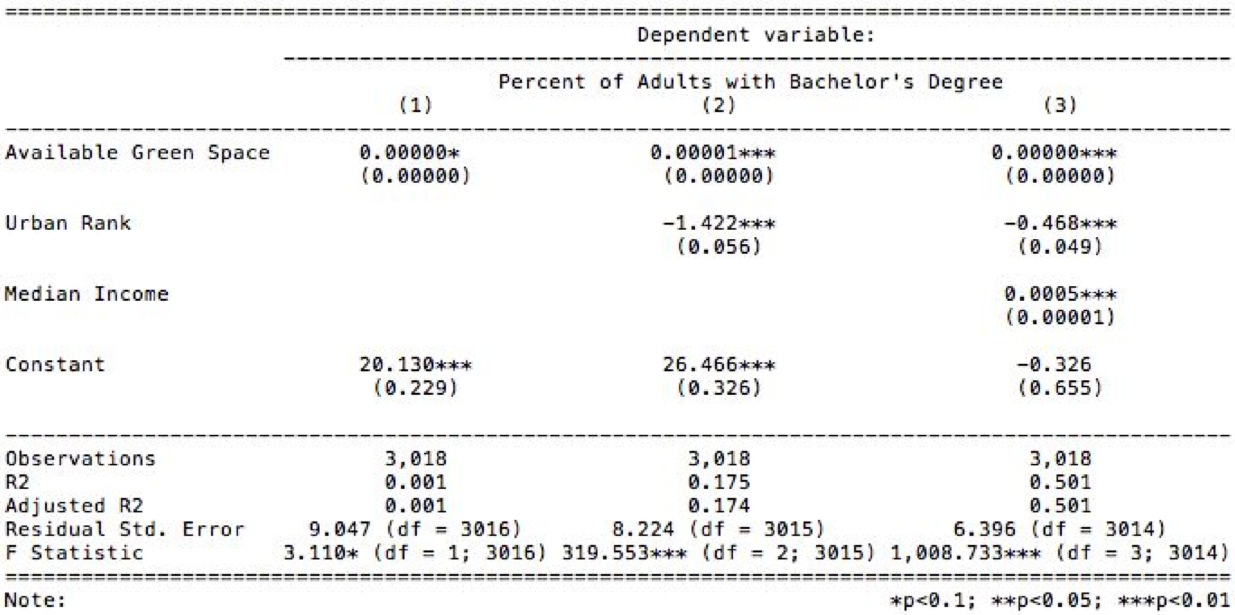


Figure 9: Education vs Available Green Space, grouped by Rural-Urban Classification

Another way we assessed environmental quality was through comparisons of available green space in each county. We omitted outliers for counties that had more that 1,000,000 ha of available green space, as this skewed our regression lines and it was not representative of the thousands of other counties. Once again we placed percent of adults with Bachelor's degrees or higher on the y-axis and available green space on the the x-axis. As mentioned several times before, it is important for us to categorize counties by their degree of urbanization. This is especially true for a variable such as available green space since less developed rural regions would naturally have more green space than urban or metropolitan areas. As shown, rural and urban regions have positive slopes indicating that counties with more available green space have more adults with a Bachelor's degree and above. Metropolitan areas, on the other hand, have a slightly negative slope unlike rural and urban regions. This is a cause for further investigation.



```
===========================================================================================
                                             Dependent variable:
                      ---------------------------------------------------------------------
                                    Percent of Adults with Bachelor's Degree
                              (1)                     (2)                      (3)
-------------------------------------------------------------------------------------------
Available Green Space      0.00000*               0.00001***               0.00000***
                          (0.00000)              (0.00000)                (0.00000)

Urban Rank                                         -1.422***                -0.468***
                                                   (0.056)                  (0.049)

Median Income                                                               0.0005***
                                                                            (0.00001)

Constant                  20.130***              26.466***                  -0.326
                           (0.229)                (0.326)                   (0.655)

-------------------------------------------------------------------------------------------
Observations                3,018                  3,018                    3,018
R2                          0.001                  0.175                    0.501
Adjusted R2                 0.001                  0.174                    0.501
Residual Std. Error    9.047 (df = 3016)      8.224 (df = 3015)        6.396 (df = 3014)
F Statistic          3.110* (df = 1; 3016) 319.553*** (df = 2; 3015) 1,008.733*** (df = 3; 3014)
===========================================================================================
Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Figure 10: Regression for Education vs Available Green Space

As shown in **Figure 10**, We ran three regressions to test the relationship between the percent of adults and amount of PM. The first regression included the two variables of interest, percent of adults with a Bachelor's degree and available green space. The second regression included Urban Rank as a control and the third included both Urban Rank and income as control variables.

The linear regression found that there is no correlation between the percent of adults with Bachelor's degrees and available green space. Based on on the results from the scatterplot, this is not what we expected to see. Perhaps a regression that examined the relationship between education and green space for each Urban Rank would be more telling than a regression that only controls for it.

<u>Geospatial Visualization (Appendix: "Data Visualization")</u>

Beyond visualizing our data using various scatter plots and observing the holistic trends they present, we also pursued visualizing our data geospatially with choropleth maps. In effect, given that we constructed a large data frame mapping relevant data to each county, the ability to see the data projected directly onto a United States map could reveal geographic trends which would otherwise be obscured if we only used plots to visualize.

We mapped our data using two core geospatial visualization methods available for R: ggplot2 spatial mapping and interactive Leaflet maps. Both methods fundamentally require a Shapefile containing current geographic boundaries for all United States counties. This county-based Shapefile was obtained from the 2015 U.S. Census Bureau's Shapefiles. Specifically in the interest of minimizing disk space and optimizing visualization rendering time, we used the simplified boundary representations from the Census Bureau's MAF/TIGER geographic database called cartographic boundary shapefiles. (See link for full technical documentation). The core structure of these shapefiles contain all the county boundaries and geographic entity codes (GEOIDs). This structure allows us to link the geographic data to our other data by mapping GEOIDs to each county's unique FIPS code identifier. The Shapefile data was obtained from the Census Bureau's website via the specialized R package *tigris* from the CRAN repository. This package directly loads the Census TIGER/Line Shapefile into R as

SpatialDataFrame objects by specifying the year of interest for the shapefile.

The ggplot2 method requires that this SpatialDataFrame be converted to a regular data frame via a special conversion package. Afterwards, county data is joined to the spatial data using a left join learned in class. Beyond this, standard ggplot functions for spatial mapping, such as geom_polygon, are used to map county data with a color gradient based on data values. We use this ggplot2 plotting method to visualize the arsenic concentrations in drinking water, as shown in **Figure 11.**

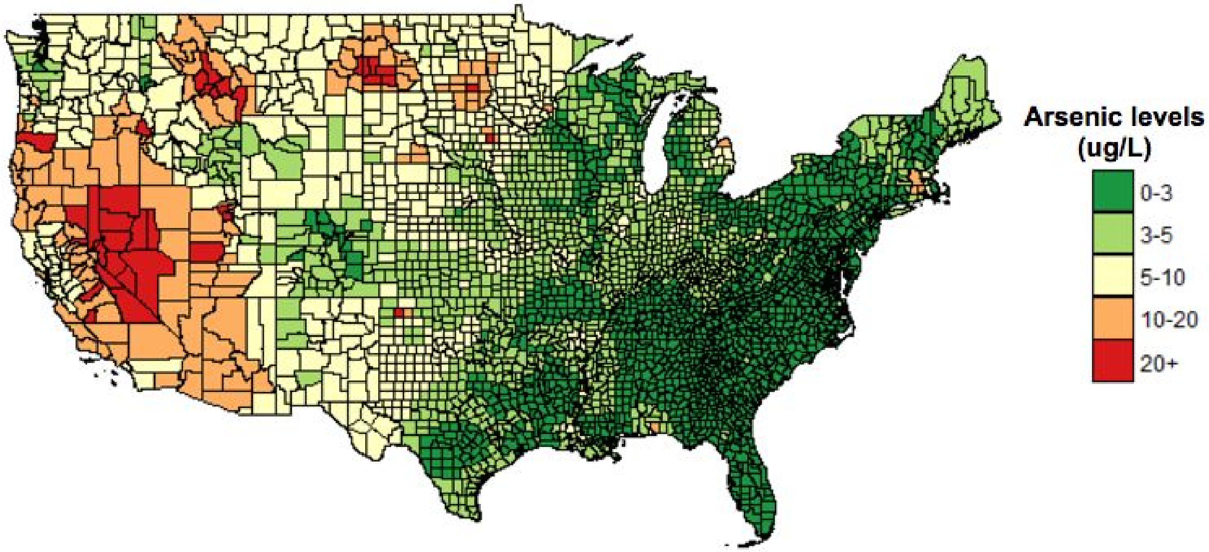


Figure 11: The ggplot map of arsenic levels in drinking water (ug/L).

There were several inherent challenges in plotting the arsenic data. When we mapped the arsenic concentrations, the gradient did not accurately represent the range of data. This was because most of the concentrations were under 20 ug/L but some points neared 50 ug/L and a few extremes ranged from 50 to over 100 ug/L. These high arsenic concentrations diluted the gradient of the lower concentrations. Before 2001, the maximum permitted arsenic concentration regulated by the EPA was 50 ug/L (USGS 2011). Therefore, we assessed that concentrations less than or equal to 50 ug/L were reasonable arsenic measurements. Any measurements above 50 ug/L were considered to be outliers and omitted from our analysis. Then, to represent the gradient more accurately, we created a new factor variable to group the arsenic concentrations into ranges. We chose the ranges 0-3, 3-5, 5-10, 10-20, and everything 20 and above. This grouping was better for the visualization of this data because it showed the areas of at-risk (5-10), high (10-20), and dangerous (20+) levels of arsenic contamination more clearly.

Now we can see that several areas throughout the US have high concentrations of arsenic, namely the West and parts of Montana, Wyoming, Idaho, North Dakota, and South Dakota. As mentioned above, arsenic pollution generally comes from pesticides, metal mining, and geothermal activity. These high arsenic regions are largely associated with mining. Large portions of California with high arsenic concentrations are also associated with agricultural zones. Counties near Yellowstone National Park also seem to have higher concentrations of arsenic. The eastern half of the US is generally clean of arsenic contamination. However, this ggplot visualization method is not ideal as the maps are inherently static and the thousands of county boundaries impede interpretation of the color gradient.

Consequently, we also pursued choropleth mapping using an R package interface of the interactive Javascript ‘Leaflet’ maps. For this method, shapefiles are directly merged with our data using a specialized left join function from the tigris package. The merged data frame is then directly inputted into the Leaflet plotting functions, allowing our data to be plotted on each county using customizable a color gradient. Leaflet is distinct from ggplot2 mapping in that Leaflet plots the map as an interactive htmlwidget. Thus, maps generated with leaflet in the R console, in markdown documents, or as Shiny widgets have interactive features such as the ability to zoom into areas dense with counties and display relevant data values when each county is clicked on the map. We utilized Leaflet to plot several variables of interest, and for the purposes of this report, we included static screenshots **(Figure 12-15).** To explore the interactive functionality of the leaflet maps, please see LeafletMaps. We tried to knit and post the Leaflet maps to Rpubs for greater ease of viewing, but it exceeded the maximum upload file size limits. These plots do not necessarily allow us to draw distinct conclusions as the scatter plots and regressions do. However, they are still a useful tool to gather spatial observations about our data and better understand the regional influence of our variables.

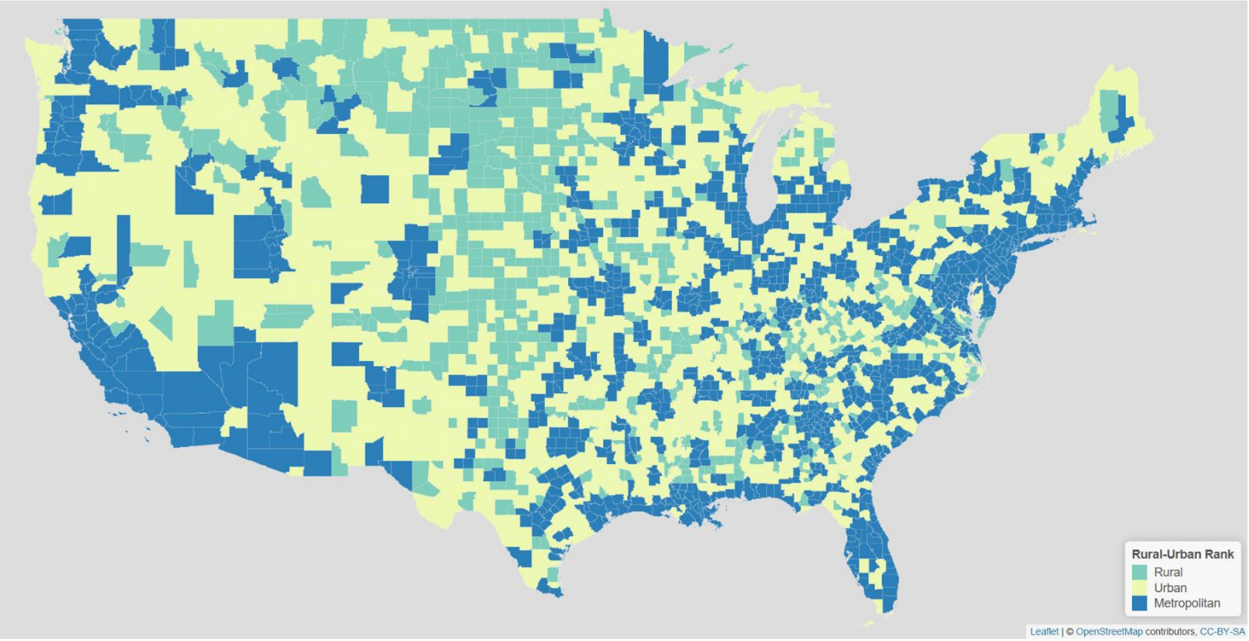


Figure 12: Leaflet plot of Rural, Urban, and Metropolitan areas as based on each county’s Rural-Urban Continuum Code.

This is the geospatial visualization of the Rural-Urban continuum codes (Urban Rank). It shows the densely populated counties surrounding large cities in the US, notable around the Eastern Seaboard and most of Southern California. The rural areas are primarily located in the agricultural areas of the midwest.

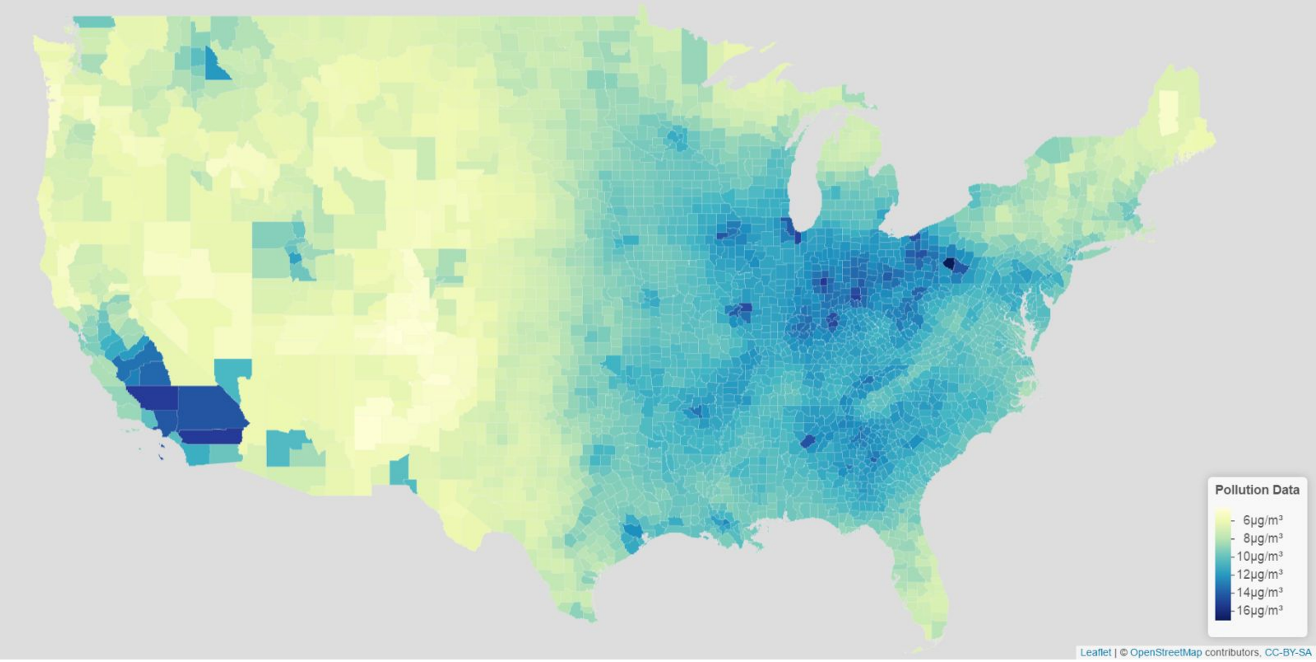


Figure 13: Static image of interactive Leaflet plot of air quality (Particulate Matter 2.5)

The Leaflet map in **Figure 13** geospatially visualizes air quality across every county in the United States. This map is notable, as it clearly indicates the regions of poor air quality, as represented by dark blue regions. The areas of particularly low air quality are widespread among metropolitan and highly industrial areas such as Los Angeles area and densely populated areas of the North East, as shown in **Figure 12**.

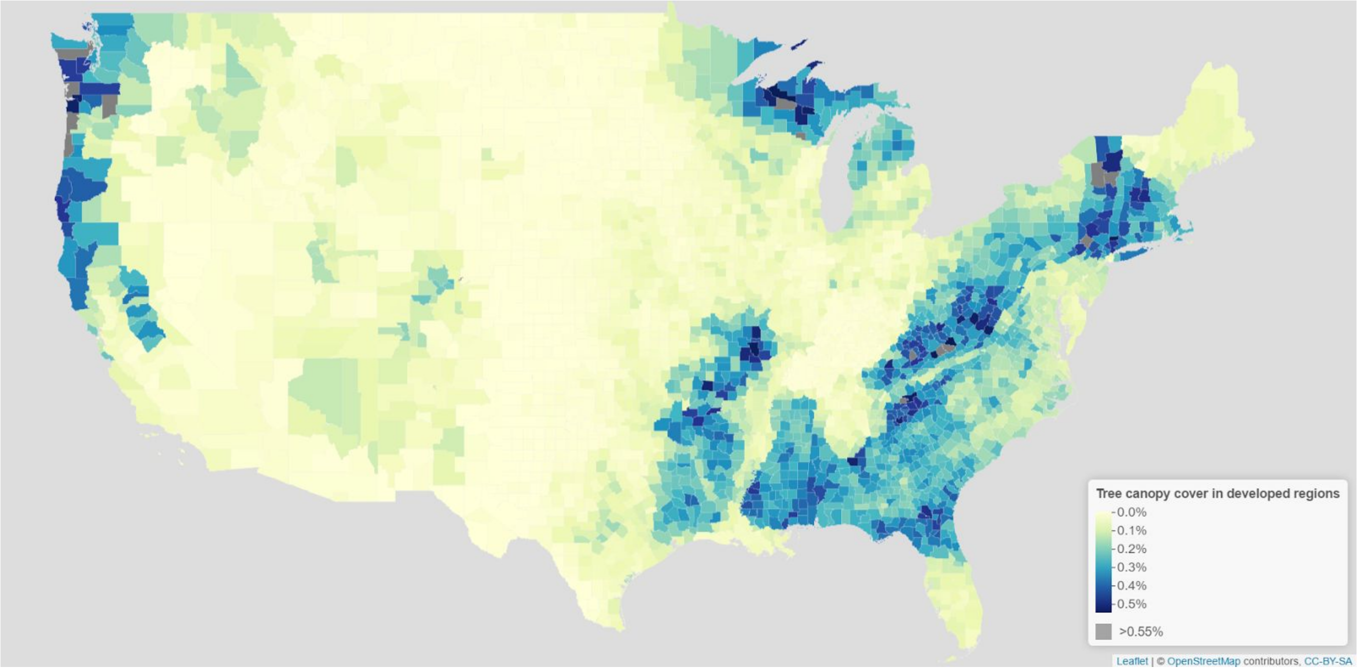

Tree canopy cover in developed regions
0.0%
0.1%
0.2%
0.3%
0.4%
0.5%
>0.55%
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 14: Static image of interactive Leaflet plot of tree canopy cover in developed regions (%).

The Leaflet map in **Figure 14** visualizes percent tree canopy cover in developed regions**.** From this map, we can clearly see a lack of tree canopy in the most metropolitan areas of Southern California, New York City, and Atlanta. Only in the Pacific Northwest, around the Great Lakes, and near the Eastern mountain range do we find urban areas with abundant tree cover.

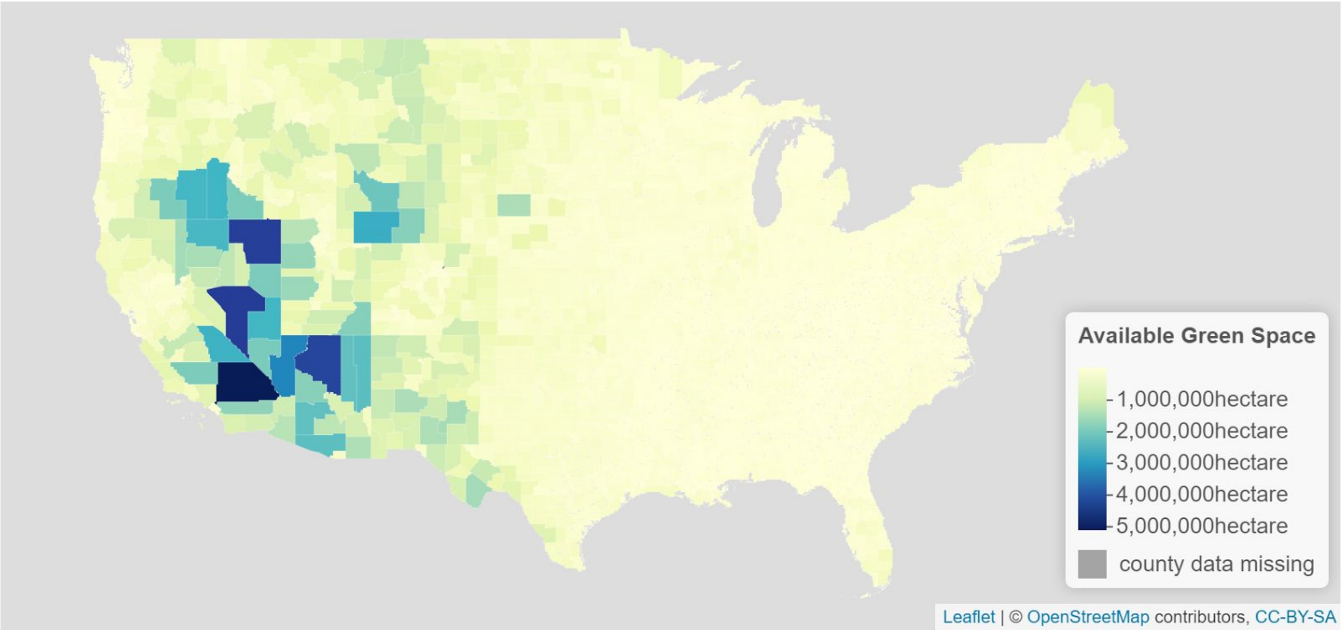


Figure 15: Static image of interactive Leaflet plot of available green space (hectare)

This map in **Figure 15** indicates availability of green space in each county (in ha). However, upon visualizing this data, it does not agree with previous maps that indicate less greenery in metropolitan areas. In fact, the plot appears to show the most available green space in the

Western desert. In effect, by graphing this data we gained a more thorough understanding of the data than would have been attained via the data descriptors and technical documentation alone.

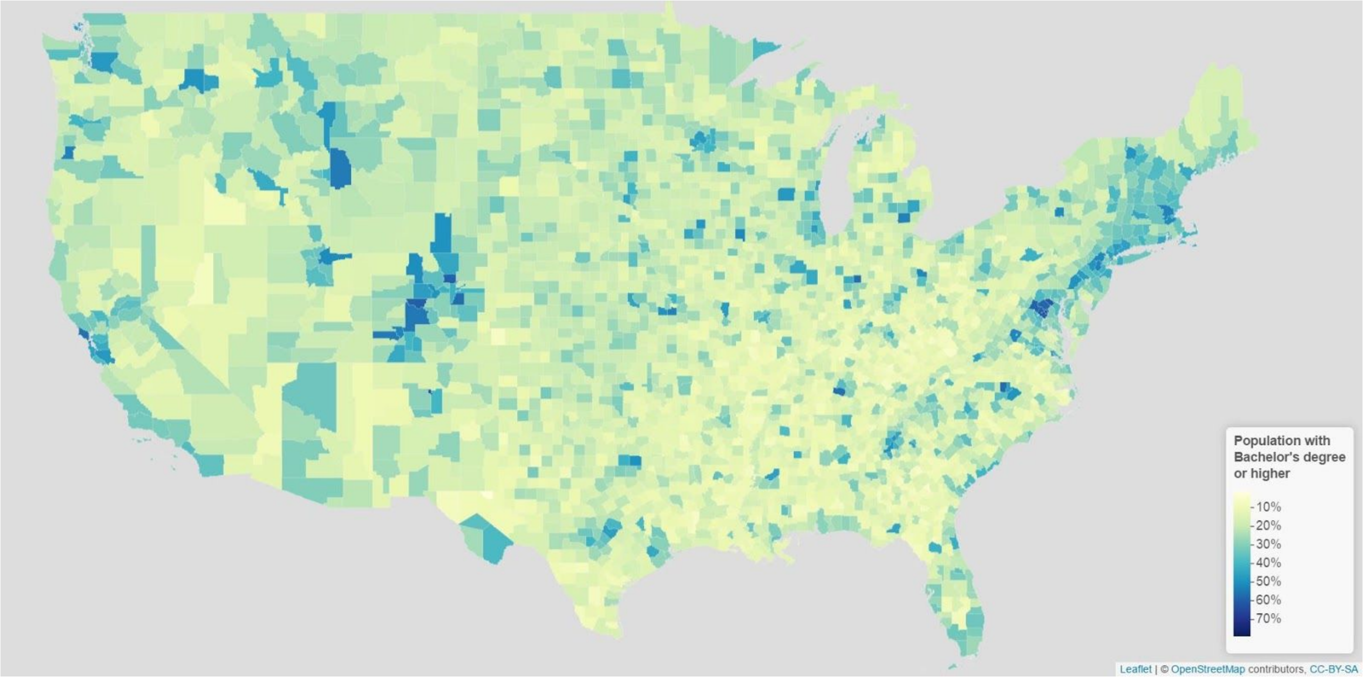

Population with Bachelor's degree or higher
10%
20%
30%
40%
50%
60%
70%
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 16: Static image of interactive Leaflet plot of population with Bachelor's degree or higher (%)

Finally, **Figure 16,** visualizes the percent of each county that has a Bachelor's degree or higher. As would be expected, the more densely populated metropolitan areas dense in business and technology such as the California Bay Area, New York City, Research Triangle, and Washington DC have the densest population of college educated citizens. However, the high percent of college educated adults in some Colorado and Wyoming counties was unexpected and deserves further investigation in a future project.

Overall, the leaflet plots were more exploratory in the context of our project, and primarily revealed interesting observations about our data in relation to geographic region. Given additional time, we would have integrated all of these Leaflet plots into a Shiny widget. This would allow us to create layer controls for our variables and allow users to directly compare several variables on one interactive map. However, given the time constraints of our project, it simply was not feasible to develop the infrastructure of the Shiny app.

Conclusion

Overall, we found statistically significant correlations between most of our environmental data and education level. However, many of the correlations were not what we had originally expected to find. For example, arsenic concentrations and education level had a generally positive correlation, which suggests that areas of high arsenic pollution also tend to have higher educational attainment. Tree canopy cover results were also unexpected; our regression analyses showed that educational attainment decreased with increased tree canopy cover. We expected a positive correlation with green space and education, but we found no significant correlation between the two. The only environmental data that had an expected outcome was the PM 2.5 concentrations and education—the regression analysis showed a negative correlation meaning that in areas with higher PM 2.5 concentrations, educational attainment is lower. However, what was particularly interesting was that the correlation became positive for tree canopy cover and education if we controlled for both Urban Rank and Income. The adjusted R^2 also increased when we controlled for both Urban Rank and Income not only for tree canopy cover, but for most other relationships. This suggests that there are likely more variables that affect the relationship between socioeconomic standing and environmental quality that we did not account for. This is a complex issue that needs to be addressed in more detail.

We only chose to examine four different environmental variables given the scope of this project and the time constraints. However, there are a myriad of variables that affect air quality, such as sulfur dioxide and ground-level ozone; water quality, such as dissolved organic matter and pH; and overall environmental health, such as proximity to landfills and mines. Not only is this issue of environmental injustice a complex problem with many influencing variables, but many of these variables are not well-monitored in underserved communities and therefore undocumented. In addition, with the current political climate, the resources to expand these services may decrease. But a robust and consistent data set for such variables are needed to fight environmental injustice throughout the country, and environmental health will likely become a much larger issue with industrial deregulation and increasing climate change.

References

Bullard, R. 2000. Dumping in Dixie: Race, Class, and Environmental Quality. 3rd Edition. Westview Press, Colorado.

Environmental Protection Agency. 2016a. Particulate matter (PM) basics. Particulate Matter (PM) Pollution. Accessed 8 May 2017. <https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#PM>

Environmental Protection Agency. 2016b. Health and environmental effects of particulate matter.

Particulate Matter (PM) Pollution. Accessed 8 May 2017. <https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm>

National Research Council. 1999. Arsenic in Drinking Water. National Academy Press, Washington D.C.

United States Geological Survey. 2011. Arsenic in groundwater of the United States. Trace Elements National Synthesis Project. Accessed 8 May 2017. <https://water.usgs.gov/nawqa/trace/arsenic>

United States Geological Survey. 2016. Contaminants found in groundwater. USGS Water Science. Accessed 8 May 2017. <https://water.usgs.gov/edu/groundwater-contaminants.html>

Ward, N. L. 2006. Improving equity and access for low-income and minority youth into institutions of higher education. Urban Education 40: 55-70.

Welch, A.H., D. R. Helsel, M.J. Focazio, and S.A Watkins. 1999. Arsenic in groundwater supplies of the United States. Arsenic Exposure and Health Effects, W.R. Chappell, C.O. Abernathy and R.L. Calderon, Eds., Elsevier Science, New York.

Wolch, J. R., J. Byrne, and J. P. Newell. 2014. Urban green space, public health, and environmental justice: The challenge of making cities “just green enough.” Landscape and Urban Planning 125: 234-244.

Appendix

~~~
------------------------------
------------------------------
DATA GATHERING
------------------------------
------------------------------
```{r include=FALSE}
library(DataComputing)
library(RCurl)
library(readxl)
library(mosaic)
library(readr)
library(tidyr)
library(dplyr)
library(stringi)
library(XML)
library(rvest)
library(rgdal)
~~~
~~~~~~
County FIPS Reference Table Data
~~~~~~
```{r}
#' Webscrapes and wrangles a table of every US county and its corresponding FIPS code and state name.
#' Source: https://en.wikipedia.org/wiki/List_of_United_States_counties_and_county_equivalents
#' @Variables: FIPS code, County, State
#' @Year: 2014
#'
#' @return data frame containing every county name and its
#' corresponding county FIPS code, state FIPS code, and state name.

GetCountyReference <- function(){
  every_us_county <-
    "https://en.wikipedia.org/wiki/List_of_United_States_counties_and_county_equivalents" %>%
    read_html() %>%
    html_nodes(xpath = '//*[@id="mw-content-text"]/table') %>%
    html_table(fill=TRUE)
  every_us_county2<-every_us_county[[2]]
  every_county<-every_us_county2%>%
    dplyr::rename(FIPSCode=INCITS, County=`County or equivalent`, State=`State or district`)%>%
    subset(select=c( `FIPSCode`, `County`, `State`))
  every_county$FIPSCode<-stri_pad_left(every_county$FIPSCode, 5, "0")
  every_county<-mutate(every_county,StateFips=substr(every_county$FIPSCode,1,2))
  return(every_county)
}
```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Educational Attainment Data
~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```{r}
#' Scrapes the excel download link for US County Education Attainment
#' Source: "https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-
data.aspx" #' (Click Educational attainment for the U.S., States, and counties, 1970-2015)
#' @Variables of interest: Rural-Urban Continuum code and Bachelors (%)
#' @Year 2013 Rural-Urban Continuum Codes, 2011-2015 Education Attainment
#'
#' @return The edu_data that contains the FIPSCode, UrbanCode (2013 Rural-Urban Continuum Code), LessThanHS
(%), HSDiploma (%), Bachelors (%), UrbanRank.

GetEduData <- function(){

  # #Use downoload link to read the data
  url <- "https://www.ers.usda.gov/webdocs/DataFiles/48747/Education.xls?v=42762"
  #tmp is a temporary file in which the data from the excel sheet is stored
  tmp <- tempfile(fileext=".xls")
  download.file(url, destfile=tmp, mode="wb")
  #read data from sheet 1 beginning at row 5
  edu_data_unclean <- read_excel(tmp, sheet = 1, skip = 4)

  #Get rid of the temporary file tmp
  unlink(tmp)

  #----Wrangle and select variables----

  #Choose the columns from edu_data_clean that has FIPS Code, 2013 Rural-Urban Continuum Code and any
columns that have information from 2011
  edu_data<-edu_data_unclean[, grep("2011|FIPS Code|2013 Rural", colnames(edu_data_unclean))]

  #Choose columns from edu_data taht has FIPSCode, 2013 Rural-Urban Continuum Code and Percent
  edu_data<-edu_data[, grep("Percent|FIPS Code|2013 Rural", colnames(edu_data))]
```

````
  #Rename FIPS Code column to FIPSCode
  edu_data<-edu_data%>%dplyr::rename(FIPSCode=`FIPS Code`)

  #Rename the columns so that they are cleaner
  colnames(edu_data)<-c('FIPSCode', 'UrbanCode', 'LessThanHS', 'HSDiploma', 'SomeCollege','Bachelors')

  edu_data$UrbanRank <- ""

  #The for loop classifies the UrbanCode (Rural-Urban Continuum Code).
  #Code 1-3 is metropolitan, 4-7 is urban and 8 - 11 is rural.
  for (i in 1:nrow(edu_data)){
     if (is.na(edu_data$UrbanCode[i])){
       edu_data$UrbanRank[i] <- NA
      } else if (edu_data$UrbanCode[i] <= 3){
        edu_data$UrbanRank[i] <- "Metropolitan"
      } else if (edu_data$UrbanCode[i] <= 7){
        edu_data$UrbanRank[i] <- "Urban"
      } else {
        edu_data$UrbanRank[i] <- "Rural"
      }
  }
  edu_data$UrbanRank <- factor(edu_data$UrbanRank, levels = c("Rural", "Urban", "Metropolitan"))
  return(edu_data)
}
```
---END Educational Attainment Data---

~~~~~~~~~~~~~~~~~~
Air Quality Data
~~~~~~~~~~~~~~~~~~
```{r}
#' Downloads and wrangles pollution data in the form of air quality measurements from the EPA.
#' Source: https://data.cdc.gov/dataset/Air-Quality-Measures-on-the-National-Environmental/cjae-szjv
#' @Variables: Annual average ambient concentrations of PM2.5 in micrograms per cubic meter
#' (based on seasonal averages and daily measurement)
#' @Year: 2010
#'
#' @return data frame containing air quality for each county in the united states

GetPollutionData <- function(){
  poll_county_unclean<-read.csv("https://data.cdc.gov/api/views/cjae-szjv/rows.csv?accessType=DOWNLOAD")
  poll_county<-poll_county_unclean %>%
    filter(MeasureType=='Average')%>%
    group_by(ReportYear)%>%
    arrange(CountyName)%>%
    arrange(StateName)%>%
    filter(MeasureId==296)%>%
    dplyr::rename(FIPSCode=CountyFips)%>%
    subset(select=c(FIPSCode, ReportYear, Value))%>%
    #Fix Shannon County change to Oglala County SD
    mutate(FIPSCode=gsub("46113","46102", FIPSCode))%>%
    filter(ReportYear==2010)%>%
    subset(select=-c(ReportYear))
  poll_county$FIPSCode<-stri_pad_left(poll_county$FIPSCode, 5, "0")
  colnames(poll_county) <- c("FIPSCode", "Pollution")
  return(poll_county)
}
```
---END Air Quality Data---

~~~~~~~~~~~~~
Income Data
~~~~~~~~~~~~~
```{r}
#' Scrapes the excel download link for US Income
#' Source: https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-
data.aspx #' (Unemployment and median household income for the U.S., States, and counties, 2007-15)
#' @Variables of interest: Income and IncomeQuartile
#' @Year 2015
#'
#' @return The income dataframe that contains FIPSCode, Income, Unemployment and IncomeQuartile

GetIncomeData <- function(){
  url <- "https://www.ers.usda.gov/webdocs/DataFiles/48747/Unemployment.xls?v=42762"
  tmp <- tempfile(fileext=".xls")
  download.file(url, destfile=tmp, mode="wb")

  income_unclean <- read_excel(tmp, sheet = 1, skip = 7)
  unlink(tmp)

  #remove any non-alphanumeric characters
````

````
  income <- income_unclean %>%
    mutate(Area_name= gsub(", ..$", "", Area_name), Area_name)%>% #remove any non-alphanumeric characters
    subset(select=c("FIPStxt","Median_Household_Income_2015","Unemployment_rate_2015"))%>%
    dplyr::rename(FIPSCode=FIPStxt)
  colnames(income) <- c("FIPSCode", "Income", "Unemployment")

  #Make a column called IncomeQuartile that separates incomes into 4 quartiles
  income$IncomeQuartile <- ntile(income$Income, 4)

  #Name the income levels
  income$IncomeQuartile[income$IncomeQuartile == 1] <- "< $40k"
  income$IncomeQuartile[income$IncomeQuartile == 2] <- "$40k - 47k"
  income$IncomeQuartile[income$IncomeQuartile == 3] <- "$47k - 54k"
  income$IncomeQuartile[income$IncomeQuartile == 4] <- "> $54k"
  income$IncomeQuartile <- base::as.factor(income$IncomeQuartile)
  income$IncomeQuartile <- factor(income$IncomeQuartile, levels = c("< $40k","$40k - 47k","$47k - 54k",">
$54k"))
  return(income)
}
```
----END Income Data-----

~~~~~~~~~~~~~~~~~~~
Forestry Data
~~~~~~~~~~~~~~~~~~~
```{r}
#' Scrapes the excel download link for each state in the U.S.
#'
#' @return A data frame containing variables: stateNames and downloadLinks (Excel file download links)

Webscrape <- function(){
  #-----------------------------------------------
  #Scrape the link to each state from the US data
  #-----------------------------------------------
  library(XML)
  library(RCurl)
  library(readxl)
  URL <- "https://www.nrs.fs.fed.us/data/urban/"
  txt <- getURLContent(URL)
  doc <- htmlParse(txt)
  #Scrape the state name
  stateNames <- xpathSApply(doc, '//ul/li/a/strong', xmlValue)
  #Scrape the state link
  stateLinks <- xpathSApply(doc, '//ul[@class="state_list"]/li/a/@href')
  baseURL <- "https://www.nrs.fs.fed.us"
  stateLinks <- paste(baseURL,as.character(stateLinks),sep="")
  #Data Frame of StateName and stateLink
  AllStates <- data.frame(stateNames, stateLinks, stringsAsFactors = FALSE)
  #Fix naming conventions standard Washington DC --> District of Columbia
  AllStates$stateNames[AllStates$stateNames == "Washington, D.C"] <- "District of Columbia"
  #-----------------------------------------------------
  #Scrape the state xls file download link from each state page
  #-----------------------------------------------------
  downloadLinks <- vector(mode="character", length=length(AllStates$stateNames))
  for (i in 1:length(stateLinks)){
    stateURL<- stateLinks[i]
    stateTxt <- getURLContent(stateURL)
    stateDoc <- htmlParse(stateTxt)
    downloadLink <- xpathSApply(stateDoc, '//ol[@id="data_options"]/li/a/@href')
    #The HTML source code is poor so need to use grepl to extract .xls from Xpath results
    length(downloadLink)
    for (j in 1:length(downloadLink)){
       if (grepl(".xls", downloadLink[j])){
        downloadLinks[i] <- downloadLink[j]
        break
      }
    }
  }
  # Add to dataframe
  AllStates$downloadLinks <- downloadLinks
  # Return only the relevant part of the database
  AllStates <- AllStates %>% select(stateNames, downloadLinks)
  return(AllStates)
}

```


```{r}
#' Creates a reference dataframe that maps county name to FIPS code based on the 2010 census,
#' with naming updates up to 2015.
#'
````

~~~
#' @param state (optional) to restrict FIPS codes to one state
#' @return A data frame containing variables: State, CountyName, and FIPS (FIPS code)

FIPS_fun <- function(state = NA){

  df <- read.table("https://www2.census.gov/geo/docs/reference/codes/files/national_county.txt", sep = ",",
col.names = c("State", "StateFIPS", "CountyFIPS", "CountyName", "ClassFIPSCode" ), colClasses =
"character", quote = "")
  #Shannon County (46-113) change to Oglala Lakota County (46-102) (Effective 2015)
  df2 <- data.frame(State = "SD", StateFIPS = "46", CountyFIPS = "102", CountyName = "Oglala Lakota
County", ClassFIPSCode = "H1")
  df <- rbind(df, df2)
  #Merge state and county into one FIPS code
  FIPS_base <- df %>% mutate(FIPS = paste(StateFIPS, CountyFIPS, sep = ""))
  #Delete Shannon County (It changed to Oglala Lakota)
  FIPS_base <- FIPS_base %>% filter(FIPS != "46113")
  #return FIPS codes of the state passed in
  if(!is.na(state)){
    FIPS_base <- FIPS_base %>% subset(State == state) %>% select(State, CountyName, FIPS)
    return(FIPS_base)
  }else{return(FIPS_base)}
}
```

```{r}
#' Extracts data from an excel file relating to the forestry variables of interest.
#' @Variables: Tree canopy (m2/person), Available green space (ha), and Tree canopy cover in developed
regions (%)
#' @param file Location of the state .xls file
#' @param stateAbbrev the state postal code for the count of interest
#' @return A data frame containing variables: State, CountyName, FIPS (FIPS code), TreeCanopy,
AvailGreenSpace, and TreeCanopyCover

ExtractStateData <- function(file, stateAbbrev){
  # -----------------------------------------------
  # Read-in relevant sheets from the excel files
  # -----------------------------------------------
  if(stateAbbrev == "DC"){
    xl_7 <- read_excel(file, sheet = "5", skip = 3)
    xl_10 <- read_excel(file, sheet = "8", skip = 4)
  }else{
    xl_7 <- read_excel(file, sheet = "7", skip = 3)
    xl_10 <- read_excel(file, sheet = "10", skip = 3)
  }
  # ------------------------------------
  # Clean and select relevant variables
  # ------------------------------------
  xl_7 <- xl_7 %>% select(c(`X__1`, `m2/person__1`, `Available green space (ha)`))
  #Units: Tree canopy Covering (m2/person), Available green space (ha)
  colnames(xl_7) <- c("CountyName","TreeCanopy", "AvailGreenSpace")
  xl_10 <- xl_10 %>% select(c(`X__1`, `Tree % h`))
  #Tree canopy cover in developed regions (%)
  colnames(xl_10) <- c("CountyName", "TreeCanopyCover")
  #Exclude the variable descriptions at the end of the sheet
  xl_10 <- na.omit(xl_10)
  # --------
  # Join
  # --------
  #Join the two excel sheets to create one datframe of county data
  joined <- full_join(xl_7, xl_10, by = "CountyName")
  #get ride of statewide summary row
  joined_clean <- joined %>% subset(CountyName != "Statewide")
  # ------------------
  # Naming Corrections:
  # -----------------
  # 1) Washington DC must be called District of Columbia to find FIPS code
  # 2) La Salle county in IL changed to LaSalle County in 2001
  # 3) Clifton Forge city is no longer a county as of 2001
  # 4) Shannon County, SD changed to Oglala Dakota in 2015
  if(stateAbbrev == "DC"){
    #joined_clean[1, "CountyName"] <- "District of Columbia"
    joined_clean$CountyName[joined_clean$CountyName == "Washington, D.C."] <- "District of Columbia"
  }else if(stateAbbrev == "IL"){
    #joined_clean[49,"CountyName"] <- "LaSalle County"
    joined_clean$CountyName[joined_clean$CountyName == "La Salle County"] <- "LaSalle County"
  }else if(stateAbbrev == "VA"){
    joined_clean <- joined_clean %>% filter(CountyName != "Clifton Forge city")
  }else if(stateAbbrev == "SD"){
    joined_clean$CountyName[joined_clean$CountyName == "Shannon County"] <- "Oglala Lakota County"
  }
  # Add FIPS codes to data
~~~

~~~
  FIPS_base <- FIPS_fun(state = stateAbbrev)
  final_df <- full_join(joined_clean, FIPS_base, by = "CountyName")
  return(final_df)
}
```

```{r}
#' Scrapes the download link for each state's forestry data and then extracts forestry data for each
county.
#' Source: https://www.nrs.fs.fed.us/data/urban/
#' @Variables: Tree canopy (m2/person), Available green space (ha), and Tree canopy cover in developed
regions (%)
#'
#' @return A data frame containing county name, FIPS code, tree canopy per person, available green space,
and
#' tree canopy cover in developed areas.

GetTreeData <- function(){
  AllStates <- Webscrape()
  #Download via temp files (no local storage on hard drive )
  for(i in 1:length(AllStates$downloadLinks)){
    url <- AllStates$downloadLinks[i]
    tmp <- tempfile(fileext=".xls")
    download.file(url,destfile=tmp, mode="wb")

    #Extract urban forestry for each state in the United States
    print(AllStates$stateNames[i])
    #Get the current state postal abbreviation
    if(AllStates$stateNames[i] == "District of Columbia"){
      stateAbbrev <- "DC"
    }else{
      stateAbbrev <- state.abb[match(AllStates$stateNames[i],state.name)]
    }

    #Extract the variables Tree canopy Covering (m2/person), Available green space (ha), and Tree canopy
cover in developed regions (%) for the current state
    state_df <- ExtractStateData(tmp, stateAbbrev)

    #Build a data frame of all the states
    if(i == 1){
      df_base <- state_df
    }else if( i == 2){
      df_full <- rbind(df_base, state_df)
    }else{
      df_full <- rbind(df_full, state_df)
    }
    unlink(tmp)
  }
  #Drop everything except for the variables of interest and FIPS code
  tree_data<-df_full[, -grep("State|Name", colnames(df_full))]
  #name to faciliate later join
  tree_data<-tree_data%>%rename(FIPSCode=FIPS)
  return(tree_data)
}
```
----END Forestry Data------

-------------
Arsenic Data
-------------
```{r}
#' Downloads Arsenic water quality data from USGS
#' Source: https://water.usgs.gov/nawqa/trace/data/index.html#ARSENIC_NOV01
#' @Year:  1973 - 2001
#' @Variables: State name,FIPS code, arsenic conentraions, lat and long
#'
#' @return data frame containing variables of interest

downloadArsenic <- function(){
  url <- "https://water.usgs.gov/nawqa/trace/data/arsenic_nov2001.xls"
  tmp <- tempfile(fileext=".xls")
  download.file(url,destfile=tmp, mode="wb")
  arsenic <- read_excel(tmp, skip=58)
  # ----DATA CLEANING----
  # get rid of non-continental US data (Alaska, Puerto Rico, Virgin Islands)
  arsenic <- arsenic[!arsenic$STATE %in% c("AK","PR","VI"),]
  # only keep columns we need (especially becuase a few of the unnecessary columns loaded in really weird)
  # columns needed: state, fips code, arsenic concentration, latitude & longitude
  arsenic <- arsenic %>%
    dplyr::select(STATE,FIPS,AS_CONC,LAT_DD,LON_DD)
  # many numeric data is in character form--change from character to numeric
~~~

~~~
  arsenic$LAT_DD <- as.numeric(arsenic$LAT_DD)
  arsenic$LON_DD <- as.numeric(arsenic$LON_DD)
  arsenic$AS_CONC <- as.numeric(arsenic$AS_CONC)
  arsenic$FIPS <- as.numeric(arsenic$FIPS)
  # take out any observations with incomplete data (eg. missing fips or arsenic)
  arsenic <- arsenic[complete.cases(arsenic[,2]),]
  return(arsenic)
}

```

```{r}
#' Interpolates arsenic data using invered distance weighted model
#' @param arsenic downloaded data frame including point data from field stations
#'
#' @return spatial data frame of interpolated data for  every ~0.25 sq mi of the US

Interpolate <- function(arsenic){
  us.map <- GetCountyShapefile()
  ## PREPPIND DATA FOR INTERPOLATION
  ## duplicate cleaned arsenic data to prevent overwriting
  arsenic_test <- arsenic
  ## assign coordinates--LON must be multiplied by negative to reflect that it is located in the W.
Hemisphere
  ## if not, this will plot on a mirrored image of the US
  arsenic_test$x <- -1*arsenic_test$LON_DD
  arsenic_test$y <- arsenic_test$LAT_DD
  ## this makes arsenic_test into a spatial data
  coordinates(arsenic_test) = ~x + y
  plot(arsenic_test)
  ## define the projection for arsenic_test to match the projection for us.map
  proj4string(arsenic_test) <- proj4string(us.map)
  ## setting the interpolation area
  ## basically the 4 most outer points of us.map in terms of longitude and latitude
  x.range <- as.numeric(c(-125, -66))  # min/max longitude of the interpolation area
  y.range <- as.numeric(c(24, 50))  # min/max latitude of the interpolation area
  ## this creates the interpolation grid
  ## it specifies the interpolatino area, and the resolution--in this case 0.1 deg. lon. x 0.1 deg. lat.
  grd <- expand.grid(x = seq(from = x.range[1], to = x.range[2], by = 0.1), y = seq(from = y.range[1], to =
y.range[2], by = 0.1))
  coordinates(grd) <- ~x + y # define spatial coordinates of grd
  proj4string(grd) <- proj4string(us.map) #with the same projections
  gridded(grd) <- TRUE
  ## visualization of grid and arsenic data
  plot(grd, cex = 1.5, col = "grey")
  points(arsenic_test, pch = 1, col = "red", cex = 1)
  ## perform the interpolation! (this will take ~20min)
  idw <- krige(AS_CONC ~ 1, arsenic_test, grd)
  return(idw)
}
```



```{r}
#' Summarizing the interpolated data to find average arsenic concentration per county
#' @param idw spatial data frame of interpolated data
#'
#' @return data frame of FIPSCode, arsenic values, and discrete values of arsenic

ArsenicbyCounty <- function(idw){
  library(raster)
  us.map <- GetCountyShapefile()
  ## convert results of interpolation from spatial pixel data to spatial raster data
  idw_raster <- raster(idw)
  ## take the mean raster value (= arsenic concentrations) within the bounds of each county defined by
us.map
  avgArsenic <- extract(idw_raster, us.map, fun = mean, sp=TRUE)
  ## convert results to data frame for data manipulation
  avgArsenic_df <- as.data.frame(avgArsenic)
  avgArsenic_df$arsenic_discrete <- NULL
  avgArsenic_df$arsenic_discrete[avgArsenic_df$var1.pred < 3] <- "0-3"
  avgArsenic_df$arsenic_discrete[avgArsenic_df$var1.pred >= 3 & avgArsenic_df$var1.pred < 5] <- "3-5"
  avgArsenic_df$arsenic_discrete[avgArsenic_df$var1.pred >= 5 & avgArsenic_df$var1.pred < 10] <- "5-10"
  avgArsenic_df$arsenic_discrete[avgArsenic_df$var1.pred >= 10 & avgArsenic_df$var1.pred < 20] <- "10-20"
  avgArsenic_df$arsenic_discrete[avgArsenic_df$var1.pred >= 20] <- "20+"
  avgArsenic_df$arsenic_discrete <- base::as.factor(avgArsenic_df$arsenic_discrete)
  avgArsenic_df$arsenic_discrete = factor(avgArsenic_df$arsenic_discrete,
levels(avgArsenic_df$arsenic_discrete) [c(1,4,5,2,3)])
  arsenic_df <- avgArsenic_df %>%
    dplyr::select(GEOID, var1.pred, arsenic_discrete)
  names(arsenic_df) <- c("FIPSCode","arsenic","arsenic_discrete")
~~~

````
  save(arsenic_df, file = "arsenic_df.RData")
  return(arsenic_df)
}
```

-----------------
Main Arsenic code
-----------------
```{r}
#' Downloads, interpolates, ans summarises Arsenic water quality data from USGS
#' Source: https://water.usgs.gov/nawqa/trace/data/index.html#ARSENIC_NOV01
#' @Year:  1973 - 2001
#' @Variables: FIPS code, arsenic conentraions, latitude, longitude
#'
#' @return data frame containing FIPS code, arsenic conentraions, and discrete arsenic range

GetArsenicData <- function(){
  arsenic <- downloadArsenic()
  idw <- Interpolate(arsenic)
  arsenic_df <- ArsenicbyCounty(idw)
  return(arsenic_df)
}
```

------END Arsenic Data---------

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
****Acquire and Join all data****
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

```{r}
#' Main code to aquire and join all data into the aggregate dataset including data for education, income,
pollution, and forestry

all_data=NULL
#Get reference table of counties to FIPS codes
every_county <- GetCountyReference()
#Acquire all data using support functions
edu_data <- GetEduData()
income_data <- GetIncomeData()
poll_data <- GetPollutionData()
tree_data <- GetTreeData()
arsenic_data <- GetArsenicData()
#Join all data to prepare for visualization
all_data<-plyr::join_all(list(every_county, edu_data, income_data, poll_data, tree_data, arsenic_data),
by='FIPSCode', type='full')
```

------------------------------
------------------------------
DATA VISUALIZATION
------------------------------
------------------------------
```{r include=FALSE}
library(DataComputing)
library(readr)
library(rgdal)
library(sp)
library(rgeos)
library(rgdal)
library(maptools)
library(dplyr)
library(leaflet)
library(scales)
library(tigris)
library(mapproj)
library(ggplot2)
```
---------------------------
Shapefile for Visualization
---------------------------
```{r}
#' Downloads and loads the county-level simplified geographic boundary shapefile from the US Census
Shapefiles.
#' Source: https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html
#' @Year: 2015
#'
#' @return spatial data frame of county-level geographic boundaries
GetCountyShapefile <- function(){
  library(tigris)
  # Download county shape file from US Census using streamlined TIGRIS package.
````

~~~
  us.map <- tigris::counties(cb = TRUE, year = 2015)
  # ----Clean county shapefile----
  # Remove Alaska(2), Hawaii(15), Puerto Rico (72), Guam (66), Virgin Islands (78), American Samoa (60)
  # Mariana Islands (69), Micronesia (64), Marshall Islands (68), Palau (70), Minor Islands (74)
  us.map <- us.map[!us.map$STATEFP %in% c("02", "15", "72", "66", "78", "60", "69",
                                          "64", "68", "70", "74"),]
  # Make sure other outling islands are removed.
  us.map <- us.map[!us.map$STATEFP %in% c("81", "84", "86", "87", "89", "71", "76",
                                          "95", "79"),]
  return(us.map)
}

```
---------------------------------------
Education (Bachelors map) Leaflet Plot
---------------------------------------
```{r, echo=FALSE}
countyShapefile <- GetCountyShapefile()
small_data <- all_data %>%
  select(FIPSCode, Bachelors)
# Merge spatial df with air quality data.
counties <- geo_join(countyShapefile, small_data, "GEOID", "FIPSCode", how = "left")
# Format popup data for leaflet map.
popup_dat <- paste(sep = "<br/>",
  "<b>County: </b>",
  counties$NAME,
  "<b>Value: </b>",
  counties$Bachelors)
#Let leaflet calculate the colors and labels for you
pal <- colorNumeric(
  palette =  "YlGnBu",
  domain = counties$Bachelors
)
# Render Map
map <- leaflet(counties, width="100%") %>% addTiles()
education_map <- map %>%
  addPolygons(stroke = TRUE, color = "white", weight = .1, smoothFactor = 0.5, opacity = 1,
    fillColor = ~pal(Bachelors), fillOpacity = 1, popup = popup_dat,
    highlight = highlightOptions(color = "#666", weight = 2, bringToFront = TRUE)) %>%
  addLegend("bottomright", pal = pal, values = ~Bachelors,
    title = "Population with<br>Bachelor's Degree<br>or Higher",
    labFormat = labelFormat(suffix = "%"), na.label = "county missing data",
    opacity = 1)

education_map

library(htmlwidgets)
saveWidget(education_map, file="education_map.html", selfcontained = FALSE)
```

--------------------------------------------
Air quality (Particulate matter) Leaflet Plot
--------------------------------------------
```{r, echo=FALSE}
countyShapefile <- GetCountyShapefile()
small_data <- all_data %>%
  select(FIPSCode, Pollution)
#Merge spatial df with air quality data.
counties <- geo_join(countyShapefile, small_data, "GEOID", "FIPSCode", how = "left")
# Format popup data for leaflet map.
popup_dat <- paste(sep = "<br/>",
  "<b>County: </b>",
  counties$NAME,
  "<b>Value: </b>",
  counties$Pollution)
#Let leaflet calculate the colors and labels for you
pal <- colorNumeric(
  palette =  "YlGnBu",
  domain = counties$Pollution
)
# Render Map
map <- leaflet(counties, width="100%") %>% addTiles()
pollution_map <- map %>%
  addPolygons(stroke = TRUE, color = "white", weight = .1, smoothFactor = 0.5, opacity = 1,
    fillColor = ~pal(Pollution), fillOpacity = 1, popup = popup_dat,
    highlight = highlightOptions(color = "#666", weight = 2, bringToFront = TRUE), group = "Pollution") %>%
  addLegend("bottomright", pal = pal, values = ~Pollution,
    title = "Pollution Data",
    labFormat = labelFormat(suffix = "µg/m³"), na.label = "county data missing",
    opacity = 1)
~~~

~~~
pollution_map

library(htmlwidgets)
saveWidget(pollution_map, file="pollution_map.html", selfcontained = FALSE)
```

-----------------------------
AvailGreenSpace Leaflet Plot
-----------------------------
```{r, echo=FALSE}
countyShapefile <- GetCountyShapefile()
small_data <- all_data %>%
  select(FIPSCode, AvailGreenSpace)
#Merge spatial df with air quality data.
counties <- geo_join(countyShapefile, small_data, "GEOID", "FIPSCode", how = "left")
# Format popup data for leaflet map.
popup_dat <- paste(sep = "<br/>",
  "<b>County: </b>",
  counties$NAME,
  "<b>Value: </b>",
  counties$AvailGreenSpace)
#Let leaflet calculate the colors and labels for you
pal <- colorNumeric(
  palette =  "YlGnBu",
  domain = counties$AvailGreenSpace
)
# Render Map
map <- leaflet(counties, width="100%") %>% addTiles()
availGreenSpace_map <- map %>%
  addPolygons(stroke = TRUE, color = "white", weight = .1, smoothFactor = 0.5, opacity = 1,
    fillColor = ~pal(AvailGreenSpace), fillOpacity = 1, popup = popup_dat,
    highlight = highlightOptions(color = "#666", weight = 2, bringToFront = TRUE)) %>%
  addLegend("bottomright", pal = pal, values = ~AvailGreenSpace,
    title = "Available Green Space",
    labFormat = labelFormat(suffix = "hectare"), na.label = "county data missing",
    opacity = 1)

availGreenSpace_map

library(htmlwidgets)
saveWidget(availGreenSpace_map, file="availGreenSpace_map.html", selfcontained = FALSE)
```

----------------------------------------------------
Tree Canopy Cover in developed regions Leaflet Plot
----------------------------------------------------
```{r, echo=FALSE}
countyShapefile <- GetCountyShapefile()
small_data <- all_data %>%
  select(FIPSCode, TreeCanopyCover)
#Filter out outlier values
small_data<-small_data%>%filter(TreeCanopyCover<.55)
#Merge spatial df with air quality data.
counties <- geo_join(countyShapefile, small_data, "GEOID", "FIPSCode", how = "left")
# Format popup data for leaflet map.
popup_dat <- paste(sep = "<br/>",
  "<b>County: </b>",
  counties$NAME,
  "<b>Value: </b>",
  counties$TreeCanopyCover)
#Let leaflet calculate the colors and labels for you
pal <- colorNumeric(
  palette =  "YlGnBu",
  domain = counties$TreeCanopyCover
)
# Render Map
map <- leaflet(counties, width="100%") %>% addTiles()
treeCanopyDeveloped_map <- map %>%
  addPolygons(stroke = TRUE, color = "white", weight = .1, smoothFactor = 0.5, opacity = 1,
    fillColor = ~pal(TreeCanopyCover), fillOpacity = 1, popup = popup_dat,
    highlight = highlightOptions(color = "#666", weight = 2, bringToFront = TRUE)) %>%
  addLegend("bottomright", pal = pal, values = ~TreeCanopyCover,
    title = "Tree canopy cover in developed regions",
    labFormat = labelFormat(suffix = "%"), na.label = ">0.55%",
    opacity = 1)
treeCanopyDeveloped_map
library(htmlwidgets)
saveWidget(treeCanopyDeveloped_map, file="treeCanopyDeveloped_map.html", selfcontained = FALSE)
```

-----------------
~~~

````
Arsenic (ggplot)
-----------------
```{r, eval = FALSE}
#Get county geographic boundaries
us.map <- GetCountyShapefile()
#Converting shapefile to dataframe for visualization
library(broom)
county_map <- tidy(us.map, region="GEOID")
small_data <- arsenic
#Merge spatial df with
counties <- left_join(county_map, arsenic, by = c("id" = "FIPSCode"))
counties %>%
  ggplot() +
    #blank county map
    geom_polygon(aes(x=long, y=lat, group=group), fill = "white", color="black", size=0.25) +
    #Arsenic data
    geom_polygon(aes(x=long, y=lat, group=group, fill = counties$arsenic_discrete), color="black",
size=0.25) +
    ggplot2::coord_map() +
    scale_fill_brewer(palette = "RdYlGn", direction = -1)  +
    labs(title="Arsenic Levels in drinking water (ug/L)") +
    theme_bw() +
    theme(plot.title=element_text(hjust=0.5),
          axis.line=element_blank(),
          axis.text.x=element_blank(),
          axis.text.y=element_blank(),
          axis.ticks=element_blank(),
          axis.title.x=element_blank(),
          axis.title.y=element_blank(),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          panel.border = element_blank(),
          panel.background = element_blank(),
          legend.title=element_blank())
```
-----------------------
UrbanRank Leaflet Plot
-----------------------
```{r, eval=FALSE, include=FALSE}
countyShapefile <- GetCountyShapefile()
small_data <- all_data %>%
  select(FIPSCode, Bachelors, UrbanRank, UrbanCode)
# Merge spatial df with air quality data.
counties <- geo_join(countyShapefile, small_data, "GEOID", "FIPSCode", how = "left")
# Format popup data for leaflet map.
popup_dat <- paste(sep = "<br/>",
  "<b>County: </b>",
  counties$NAME,
  "<b>Value: </b>",
  counties$UrbanCode)
#Let leaflet calculate the colors and labels
pal<- colorFactor(
  palette = "YlGnBu", levels = factor(counties$UrbanRank, levels =c("Metropolitan", "Urban", "Rural",
ordered = TRUE))
)
# Render map
map <- leaflet(counties) %>% addTiles()
urbanrank_map <- map %>%
  addPolygons(stroke = TRUE, color = "white", weight = .1, smoothFactor = 0.5, opacity = 1,
    fillColor = ~pal(UrbanRank), fillOpacity = 1, popup = popup_dat,
    highlight = highlightOptions(color = "#666", weight = 2, bringToFront = TRUE)) %>%
  addLegend("bottomright", pal = pal, values = ~UrbanRank,
    title = "Rural-Urban Rank",
    labFormat = labelFormat(), na.label = "county missing data",
    opacity = 1)

urbanrank_map

library(htmlwidgets)
saveWidget(urbanrank_map, file="urbanrank_map.html", selfcontained = FALSE)
```
~~~~~~~~~~~~~~
Scatter Plots
~~~~~~~~~~~~~~
```{r}
library(ggplot2)
#Plots percent of adults with a Bachelor's degree vs. amount of PM 2.5 (pollution), facetted by income
bach_vs_pollution_lm_income <- all_data %>%
  na.omit() %>%
  ggplot(aes(Pollution, Bachelors)) +
  geom_point(aes(color = IncomeQuartile), alpha = 0.7, size = 2, show.legend = FALSE) +
````
````

~~~
  scale_color_manual(values = c("#a1d99b","#41ab5d","#238b45","#005a32")) +
  geom_smooth(method = "lm",  fill = "black", col = "black") +
  facet_grid(~ IncomeQuartile) +
  labs(x="Concentration of PM 2.5 (ug/m^3)", y="% Adults with Bachelor's Degrees") +
  xlim(5, 15) +
  theme_bw()

bach_vs_pollution_lm_income

#Plots percent of adults with a Bachelor's degree vs. percent of tree canopy cover, facetted by income
bach_vs_tree_lm_income <- all_data %>%
  na.omit() %>%
  filter(TreeCanopyCover<.55) %>%  #Filter out outlier values
  ggplot(aes(TreeCanopyCover, Bachelors)) +
  geom_point(aes(color = IncomeQuartile), alpha = 0.7, size = 2, show.legend = FALSE) +
  scale_color_manual(values = c("#a1d99b","#41ab5d","#238b45","#005a32")) +
  geom_smooth(method = "lm",  fill = "black", col = "black") +
  facet_grid(. ~ IncomeQuartile) +
  labs(x="% Tree Canopy Cover", y="% Adults with Bachelor's Degrees") +
  theme_bw()

bach_vs_tree_lm_income
#Plots percent of adults with a Bachelor's degree vs. amount of PM 2.5 (pollution), facetted by rural-urban
classification
bach_vs_pollution_lm <- all_data %>%
  na.omit() %>%
  ggplot(aes(Pollution, Bachelors)) +
  geom_point(aes(color = UrbanRank), alpha = 0.7, size = 2, show.legend = FALSE) +
   scale_color_manual(values = c("#8dd3c7","#80b1d3","#bc80bd")) +
  geom_smooth(method = "lm", fill = "black", col = "black") +
  facet_grid(~ UrbanRank) +
  labs(x="Concentration of PM 2.5 (ug/m^3)", y="% Adults with Bachelor's Degrees") +
  theme_bw()

bach_vs_pollution_lm

#Plots percent of adults with a Bachelor's degree vs. percent of tree canopy cover, facetted by rural-urban
classification
bach_vs_tree_lm <- all_data %>%
  na.omit() %>%
  filter(TreeCanopyCover<.55) %>%    #Filter out outlier values
  ggplot(aes(TreeCanopyCover, Bachelors)) +
  geom_point(aes(color = UrbanRank), alpha = 0.7, size = 2, show.legend = FALSE) +
  scale_color_manual(values = c("#8dd3c7","#80b1d3","#bc80bd")) +
  geom_smooth(method = "lm",  fill = "black", col = "black") +
  facet_grid(. ~ UrbanRank) +
  labs(x="% Tree Canopy Cover", y="% Adults with Bachelor's Degrees") +
  theme_bw()

bach_vs_tree_lm

#Plots Arsenic concentration vs. Bachelor's degree attainment, facetted by urban rank.
bach_vs_arsenic_lm <- all_data %>%
  na.omit() %>%
  ggplot(aes(arsenic, Bachelors)) +
  geom_point(aes(color = UrbanRank), alpha = 0.3, size = 2) +
  scale_color_manual(values = c("#6baed6","#4292c6","#2171b5")) +
  geom_smooth(method = "lm") +
  facet_grid(~ UrbanRank) +
  labs(x="Arsenic in Drinking Water (ug/L)", y="% Adults with Bachelor's Degrees") +
  theme_bw()

bach_vs_arsenic_lm

#Plots percent of adults with a Bachelor's degree vs. available green space, facetted by rural-urban
classification
bach_vs_green_lm <- all_data %>%
  na.omit() %>%
  filter(AvailGreenSpace < 10^6) %>% #filter outlier values
  mutate(AvailGreenSpace = AvailGreenSpace/100000) %>%
  ggplot(aes(AvailGreenSpace, Bachelors)) +
  geom_point(aes(color = UrbanRank), alpha = 0.7, size = 2, show.legend = FALSE) +
  scale_color_manual(values = c("#8dd3c7","#80b1d3","#bc80bd")) +
  geom_smooth(method = "lm",  fill = "black", col = "black") +
  facet_grid(. ~ UrbanRank) +
  labs(x="Available Green Space (100,000 ha)", y="% Adults with Bachelor's Degrees") +
  theme_bw()

bach_vs_green_lm
```
~~~

~~~
------------------
------------------
Linear Regressions
------------------
------------------
```{r eval = FALSE}
library(statisticalModeling)
library(rpart)
library(rpart.plot)
library(stats)
library(mosaicData)

#Tree/education regression
et1<-lm(Bachelors~TreeCanopyCover, data=all_data)
et2<-lm(Bachelors~TreeCanopyCover + UrbanCode, data=all_data)
et3<-lm(Bachelors~TreeCanopyCover + UrbanCode + Income, data=all_data)
stargazer::stargazer(et1, et2, et3, type="text",
 dep.var.labels=c("Percent of Adults with Bachelor's Deegree"),
 covariate.labels=c("Tree Canopy Cover","Urban Rank", "Median Income"), out="tree_edu.txt")

#pollution/education regressions
ep1<-lm(Bachelors~Pollution, data=all_data)
ep2<-lm(Bachelors~Pollution + UrbanCode, data=all_data)
ep3<-lm(Bachelors~Pollution + UrbanCode + Income, data=all_data)
stargazer::stargazer(ep1, ep2, type="text",
 dep.var.labels=c("Percent of Adults with Bachelor's Deegree"),
 covariate.labels=c("Particulate Matter","Urban Rank", "Median Income"), out="polution_edu.txt")

#green space/education regression
gs1<-lm(Bachelors~AvailGreenSpace, data=filter(all_data, AvailGreenSpace < 10^6)%>%na.omit(all_data))
gs2<-lm(Bachelors~AvailGreenSpace + UrbanCode, data=filter(all_data, AvailGreenSpace <
10^6)%>%na.omit(all_data))
gs3<-lm(Bachelors~AvailGreenSpace + UrbanCode + Income, data=filter(all_data, AvailGreenSpace <
10^6)%>%na.omit(all_data))
stargazer::stargazer(gs1, gs2, gs3, type="text",
 dep.var.labels=c("Percent of Adults with Bachelor's Degree"),
 covariate.labels=c("Available Green Space","Urban Rank", "Median Income"), out="green_edu.txt")

#Arsenic/education reg
ea1<-lm(Bachelors~arsenic, data=all_data)
ea2<-lm(Bachelors~arsenic +UrbanRank, data=all_data)
ea3<-lm(Bachelors~arsenic +Income, data=all_data)
stargazer::stargazer(ea1, ea2, ea3, type="text",
 dep.var.labels=c("Percentage of Adults with Bachelor's Degrees"),
 covariate.labels=c("Arsenic Level","Urban/Rural Classification","Median Income"), out="arsenic_edu.txt")
```
~~~